THE UNIVERSITY *of* EDINBURGH

An Evaluation of the Effectiveness and Predictive Validity of English Language Assessment in Two Colleges of Applied Sciences in Oman

Fatma Said Mohammed Al Hajri

Thesis

Submitted to the University of Edinburgh for the degree of Doctor of Philosophy

College of Humanity and Social Science
The Moray House School of Education
2013

Declaration

I declare that the work presented in this document is the original work of the author and that it has not been submitted for any other degree or professional qualification.

Fatma Al Hajri

Signature…….

Date……..

Acknowledgements

For the past three years, many people offered me unconditional love, support and encouragement without which this work would not have been completed. I am in debt to these people.

First, I would like to say thank you to my supervisors, Mr. Brian Parkinson and Dr. Aileen Irvine for their academic guidance and advice. Brian, thank you for your persistent work and effort in leading me through the process of conducting this research and writing up the thesis, your supervision is invaluable.

Second, I would like to thank my friends and officemates for always being around when I needed them. Rong, Han, Anna, Julie, Katerina and Siti, I am very lucky to have known you on this journey, you are my small family at Moray House. Marguerite, thank you for being such a wonderful friend and mentor. Thank you for showing interest on this study and taking the time to read it.

Third, I would like to express my deepest gratitude to my family for their endless love and support. Dad, no words can express my appreciation for what you have done for me and my siblings. Mum, thank you for your sweet love and care. Brothers and sisters, thank you for your encouragement. I would like, also, to extend my gratitude to my uncle Amer, cousins and family in law; I am blessed to have you as my extended family.

Last but not least, a special thank you to my husband without whom this work would not have been completed. Khalid, thank you for your sacrifices, patience and support. Your love inspired me in every way.

This work is dedicated to my daughter Mariam who joined me in the second year of this journey. Mariam, you are the joy of my life.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

AES	Academic English Skills (Course)
CS	Communication Studies
CA	Continuous Assessment
CAS	Colleges of Applied Sciences
CR	Criterion- Referenced
EAP	English for Academic Purposes
GES	General English Skills
GPA	Grade Point Average
FP	Foundation Programme
FY	First Year
HEI	Higher Education Institutions
IBA	International Business Administration
IELTS	International English Language Test System
IT	Information Technology
NR	Norm-Referenced
NZTEC	New Zealand Tertiary Education Consortium
OAAA	Oman Academic Accreditation Authority
PINZ	Polytechnics International NewZealand
TOEFL	Test of English as a Foreign Language

Abstract of Thesis

This thesis investigates the effectiveness of English language assessment in the Foundation Programme (FP) and its predictive validity for academic achievement in the First Year (FY) at two Colleges of Applied Sciences (CAS) in Oman.

The objectives of this study are threefold:

- Identify how well the FP assessment has met its stated and unstated objectives and evaluate its intended and unintended outcomes using impact evaluation approaches.
- Study the predictive validity of FP assessment and analyse the linguistic needs of FY academic courses and assessment.
- Investigate how FP assessment and its impact are perceived by students and teachers.

The research design was influenced by Messick's (1989; 1994; 1996) unitary concept of validity, by Norris (2006; 2008; 2009) views on validity evaluation and by Owen's (2007) ideas on impact evaluation. The study was conducted in two phases using five different methods: questionnaires, focus groups, interviews, document analysis and a correlational study. In the first phase, 184 students completed a questionnaire and 106 of these participated in 12 focus groups, whilst 27 teachers completed a different questionnaire and 19 of these were interviewed. The aim of this phase was to explore the perceptions of the students and teachers on the FP assessment instruments in terms of their validity and reliability, structure, and political and social impact. The findings indicated a general positive perception of the instruments, though more so for the Academic English Skills course (AES) than the General English Skills course (GES). There were also calls for increasing the quantity and quality of the assessment instruments. The political impact of the English language FP assessment was strongly felt by the participants.

In the second phase, 176 students completed a questionnaire and 83 of them participated in 15 focus groups; 29 teachers completed a different questionnaire and of these 23 teachers were interviewed. The main focus was on students and teachers' perceptions of FP assessment, and how language accuracy should be considered in

marking academic written courses. One finding was that most students in FY tended to face difficulties not only in English but also in what could be called ‘study skills’; some of these were attributed to the leniency of FP assessment exit criteria.

Throughout the two phases, 118 documents on FP assessment at CAS were thematically analysed. The objective was to understand the official procedures prescribed for writing and using assessment instruments in FP and compare them against actual test papers and classroom practices. The findings revealed the use of norm-referenced assessment instead of criterion referenced, incompatibility between what was assessed and what was taught, inconsistency in using assessment criteria and in the unhelpful verbatim replication of national assessment standards.

The predictive validity studies generally found a low overall correlation between students’ scores in English language assessment instruments and their scores in academic courses. The findings of this study are in line with most but not all previous studies. The strength of predictive validity was dependent on a number of variables especially the students’ specializations, and their self-evaluations of their own English language levels. Some recommendations are offered for the reform of entry requirements of the Omani higher education.

Chapter1: Introduction to English Language Education and Assessment in Omani Higher Education

1.1. Introduction

On 27th February 2011, about 200 young Omanis protested in the city of Sohar, one of the most important cities in Oman. The protest turned violent and a protester, a first year student at a college of technology, was killed by police. After this incident other protests started in almost all Omani regions. University students constituted the biggest group in these demonstrations. The Colleges of Applied Sciences (CAS) witnessed similar student strikes, demonstrations and sometimes vandalism; and English language assessment repeatedly emerged as an issue.

On 14th March 2011, most of the students at the College of Applied Sciences in Sur started a demonstration replicating concurrent demonstrations in other CAS campuses in Salalah, Al-Rustaq, Ibri, Sohar and Nizwa. The author was at Sur College collecting data when a demonstration took place. One of the main demands was to change aspects of the assessment system at the colleges (a copy of a flyer distributed in this demonstration is presented in appendix 1.1). Some of the assessment related demands as stated in the students' flyers, letters and the Omani Gazette were as follows:

- to eliminate norm-referenced assessment and award grades based on student achievements[1] instead
- to allow students who had failed the Foundation Programme[2] (FP) the previous year, as well as those who had been expelled for academic under-achievement, to retake the final exam.

[1] Students referred to norm-referenced assessment as "curve-based" assessment in their flyers.
[2] The Foundation Programme (FP) is a pre-university programme that consists of twenty hours of English language instruction, three to four hours of mathematics and two hours of computer skills courses in each semester.

- to extend the time allocated for English language examination sessions to allow students to revise their answers
- to ensure teachers' objectivity in marking.

These demands show students' concerns and recognition of the power innate in the assessment mechanisms used at CAS. Spolsky, following Foucault, claims that an examination is "a mechanism linking power and knowledge" and a "ritualized ceremony that required the subjects to be seen, and transformed them into objects under control" (1995, p.15). CAS students rebelled against this power. On the 26th March 2011 the Ministry of Higher Education responded to their demands and announced acceptance of some and consideration of others. One of the approved demands was granting the students who had failed the FP assessment in the previous academic year one more semester in which to attempt to pass it. The way in which these demonstrations affected this study is clarified in Section 4.3.2.

The present study started before these demonstrations, but it was based on similar concerns. I was as an English language teacher in one of the colleges prior to embarking on this study. The role students' proficiency in English language played in their academic achievement had always been an issue of debate amongst teachers and administrators. Some teachers of academic courses attributed students' under-achievement to their inadequate English language skills, assuming a positive correlation between students' proficiency in the medium of instruction (i.e., English) and their academic achievement. Some English language teachers agreed with this line of argument, while others seemed to believe that students' proficiency in the English language was only one of several factors that affected academic achievement. I also held the position of Deputy Director of the English language programme at CAS for almost two years, during which time this issue was continually discussed with regard to the best cut-off point[3] or proficiency level in English for the First Year[4] (FY). Some directors

[3] The cut-off point refers to the lowest score obtained in the FP assessment and acceptable to qualify students to start their academic study in the First Year.
[4] The First Year is the academic year after the Foundation Programme and marks the beginning of the official academic study at CAS.

seemed to believe that attaining higher levels in English was a major factor in better academic achievement, and attributed the high failure rates of FY students to their inadequate English language abilities, arguing that the cut-off point should be raised. The need to explore the role played by students' language proficiency in academic achievement at CAS and the need to investigate the effectiveness of the FP assessment structure and instruments were the driving forces for this study.

In studying FP assessment, the term "effectiveness" will be used to mean a judgement of worth of the FP assessment instruments and how well they meet the purposes, objectives and outcomes they were intended to meet and whether they were implemented as planned. The effectiveness of FP assessment will be studied using approaches adopted from impact evaluation including objective-based, needs-based, and goal-free approaches (for detailed explanation of these approaches, see Section 3.3.4).

In this chapter, the theoretical and contextual background of the study is presented. It starts by discussing globalisation as a catalyst in the spread of English as the medium of instruction in the higher education of non-English language speaking countries. It identifies the role of English language in the globalised world as a gatekeeper to higher education and the labour market, and it presents its role in promoting higher education and internationalising its academic programmes. After that, the discussion is narrowed to focus on the impact of globalisation on Omani higher education, especially on English language education. This is followed by background information about the focus of the study and its context. The last section of this chapter covers the rationale, objectives, and research questions and gives an outline of the whole study.

1.2. Globalisation Impact on Higher Education

Altbach and Knight (2007) argue that globalisation and internationalisation, though sometimes used in similar ways, entail different ideologies. The role of globalisation in higher education is seen as "the economic, political, and social forces pushing the 21st century higher education toward greater international involvement" (p.290), whereas

internationalisation has been described as “a two-way street; students move largely from south to north … and serves important needs in the developing world” (p. 291). The impact of globalisation on education is often considered indirect. For example, Dale (1999) asserts that “absolutely central to arguments about the effects of globalisation on public services like education is that those effects are largely indirect; that is to say, they are mediated through the effect of globalization on the discretion and direction of nation states” (p.2).

Despite this indirectness, the economic imprint of globalisation on higher education is undeniable. Within the concept of “marketization of education”, Ball (1998) identifies five ideologies that have led the educational reforms stimulated by globalisation: (a) neo-liberalism, (b) new institutional economics, (c) performativity (i.e., indirect steering through “target setting” and “accountability”), (d) public choice theory, and (e) new managerialism. It is suggested that these ideologies have been disseminated in four ways: (a) policy borrowing, (b) the movement of graduates, (c) policy entrepreneurs, and (d) sponsorship (for a more detailed discussion on this see Dale, 1999). The commercialisation of education not only entails using curriculum, textbooks and policies as commodities but also includes trading with accreditation programmes and selling accreditation services around the world. Altbach and Knight (2007, p.301) observe that “the accreditation process is becoming internationalised and commercialised … national and international accreditation agencies now work in many countries”. Thus, education has been increasingly turned into an international commodity that not only generates profit but also gradually transforms local education systems so that “one size fits all” (Donn & Al Manthri, 2010).

Neo-liberalism or the “ideologies of the market” (Ball, 1998, p.122) has long been identified as one of the factors that has driven education reforms in the westernised post-industrialised countries. It has been described as “a political project for facilitating the re-constructing and re-scaling of social relations in accordance with the demand of an unrestrained global capitalism" (Fairclough, 2003, p.4). Neo-liberalism has been shown

to influence education in general and assessment in particular for more than a decade in the wider education literature. The reforms in education and consequently assessment policies are linked by several authors to the growing global concept of a 'knowledge-based economy' (Grek, 2009; Al Rahbi, 2008). Grek claims that education "has been reframed as central to national economic competitiveness within an economistic human capital framework and linked to an emerging knowledge economy" (p.24). She explores the impact of the Programme for International Student Assessment (PISA), and claims that "[PISA], through its direct impact on national education systems in Europe and beyond, has become an indirect, but nonetheless influential tool of the new political technology of governing the European education space by numbers" (p.23, 2009). In capitalist societies, the influence of neo-liberalism in education is expressed through establishing policies to enhance the global 'knowledge economy' or 'capital' but sometimes also enforced by less obviously capitalist programmes such as PISA.

1.2.1. Globalisation and English Language Education

Nonetheless, the role of neo-liberalism in the spread of English language teaching and assessment reforms has only recently been considered in the English language assessment literature. Though this phrase was not used in his discussion of the spread of English language and related policies, Alderson (2009) discusses previous authors' views of English language education as a "neo-colonist enterprise" that had served the interests of the countries of its origin such as the UK, USA and Australia and other commercial hegemonies. His discussion focuses on the micro-politics of language education as opposed to macro-politics, but these are intertwined. Macro-politics is explained as being concerned with issues of power and how it is used by countries or organisations to accomplish specific goals, and micro-politics with individuals and their use of power. The contributors to Alderson (2009) explore the role of micro-politics in language education in different countries using, many extended individual narratives of behaviours "with and around power". Several other authors have discussed the history and politics of the English language, its spread in the world today, and the influence this

has had on education (see, for example, Pennycook, 1994; 1999; Phillipson, 1992). From such arguments, it is clear that the role of English language assessment in higher education is very complicated and affected by multiple factors other than those directly related to educational assessment, such as colonisation, globalisation, internationalisation and neo-liberalism. The following section describes the context of higher education and discusses its reforms and policies with reference to globalisation. It also gives a general account of the role of English language in higher education.

1.2.2. English Language Assessment and Access to Higher Education

Proficiency in English language and how it is measured have become central issues in higher education research as the English language is increasingly used as a medium of instruction and a criterion for admission to education. In a review of a number of articles about language policies in Asian higher education, Ross (2008, p.8) states that "a commonly accepted assumption is that a foreign language learned in the context of formal schooling yields suitable subject matter for making high-stakes inferences about qualifications for admissions or employment"; he explains that there is a growing trend to use test scores in determining access to higher education, and that proficiency in the English language has also become a dominant criterion for success in the labour market. Similarly, Altbach and Knight (2007) assert that the increasing trend of using English as a medium of instruction in many higher education institutes has been stimulated by commercial factors; they state that "the growing use of English as a medium of research and instruction, especially at the graduate level, may stimulate interest in international programmes offered by universities in English" (p.303).

In Germany, for example, Erling and Hilgendrof (2006) describe the current role played by English language in German higher education and attribute this role to economic forces. They state that:

> in an effort to internationalise the curriculum and become more competitive in the global market for students, German institutions have chosen the English language as an important strategy for achieving their goals (p.287).

They expressed a concern, however, that staff and students' possible inadequate ability to discuss advanced topics in English could affect the quality of education (Erling and Hilgendrof, 2006). With an increased number of students joining higher education (Altbach & Knight, 2007), it is important to understand how the English language has become a gatekeeper and has been turned into a commercial tool for marketing programmes, attracting international students, promoting graduates or controlling access to the labour market. The next section discusses the impact of globalisation on Omani higher education and the role played by the English language assessment in admission to higher education.

1.3. Higher Education in Oman

Omani higher education has both private and public sectors. *Public* higher education commenced in 1986 and continued to be the sole form of higher education until 1996 when the first *private* colleges were established. The private institutions are either affilliated with internationally recognised universities (e.g., Sohar university is affiliated with Queensland University, Australia), campuses of cross-border universities (e.g., the German University of Technology (GUTEC)), or independent institutions offering locally developed programmes with no overt association with foreign higher education providers (e.g., The University of Nizwa).

Private higher education has been seen as a solution for the tension between the increasing demand for higher education and the limited number of scholarships available in the public system. In Oman the number of high school graduates increases every year, as does the number of applicants to higher education. In 2009, for example, there were 80,000 students in higher education (Al Shemli, 2009, p.2), but in the same year, only 14,168 high school graduates were offered full or partial scholarships (see Table 1.1). It is projected that generally the demand for international higher education in Middle East countries will continue to rise as well as the demand for "education providers across national borders" (Altbach & Knight, 2007, p.295). The government has encouraged the private sector to become involved in higher education and has offered attractive subsidy

packages; as a result, the number of private higher education institutions rose rapidly (see Al-Lamki, 2006, for a detailed description of the rise in private higher education institutions). Several new campuses of western universities have recently opened or are planned in the near future in various Middle East countries including Oman (ibid). In 2009, there were 24 private institutions offering higher education programmes with the prospect of more institutions joining this sector (Al Shemli, 2009).

The rise of the private sector in Oman higher education is also evident in the number of applications received every year. It can be seen from Table 1.1 that the number of applications to public higher education institutions decreased slightly in 2009/2010 compared to 2008/2009, despite the increased number of registered high school graduates, but there was an increase in applications to private higher education institutions (Higher Education Admission Centre, 2010). It should be noted here that this decrease in the number of applicants to public higher education institutions probably does not imply less interest in the free scholarships the public higher education offers; rather it could indicate the applicants' despair of accessing public higher education due to the low percentage of admitted applicants: 30.3% in 2009 and 31.3% in 2010.

Table 1.1. Number of Applicants and Admitted Students to *Public* Higher Education Institutions in 2008/2009 and 2009/2010

Academic Year	**2008/2009**			**2009/2010**		
Gender	*Male*	*Female*	*Total*	*Male*	*Female*	*Total*
Registered	24654	24930	49584	24687	25678	50365
Applicants	22905	23782	46687	21422	23767	45189
Admitted	8260	5908	14168	8320	5988	14308
% Enrolled Applicants	36.1%	24.8%	30.3%	38%	25.2%	31.3%

Source: Higher Education Admission Centre (2010)

However, private higher education in Oman still faces several obstacles: it struggles to make a profit in its first stages, which has sometimes led to the lowering of quality of the education offered to reduce costs (Donn & Al-Manthri, 2010). Equally, the sudden increase in private higher education, and priortising of profit in this fairly new sector has raised concerns about its quality (Al Shemli, 2009). Donn and Al-Manthri discuss the tendency of some private institutions in Oman to hire part-time academics with lower qualifications; they urge developing high-quality private education as this would be a

"key force in the global politico-economy" (2010, p.112). Some authors have commented on the vital role of monitoring and assuring the quality of higher education undertaken by the Omani Academic Accreditation Authority (OAAA) (Al Bandary, 2005). However, the OAAA, being recently established, has concentrated all its attention and resources on accrediting institutions, not academic programmes. Currently, only institutional systems such as "governance and management", "academic support services" and "staff research and consultancy" are reviewed for quality assurance purposes. However, it is the academic programmes that are in urgent need of revision, review and accreditation.

Another problem that has resulted from increasing the number of private institutions is the duplication of specialist programmes such as Information Technology (IT) and International Business Administration (IBA) in both private and public institutions (Al-Lamki, 2006). This has been ascribed either to the lack of needs analysis or to building higher education solely on the expected needs of the labour market (Al-Lamki, 2006; Donn & Al-Manthri, 2010). Meanwhile, these institutions fill the market with graduates of dubious quality and redundant academic degrees. All in all, private higher education in Oman is still struggling in terms of quality, target and profit.

1.3.1. Globalisation Impact on Oman Higher Education

In 2002, Oman signed the General Agreement on Trade in Services (GATS) and became part of the global competitive market, and consquently had to upgrade the quality of its labour force to survive the competition. Donn and Issan (2007) note the impact of signing this agreement on Oman higher education: "Oman has become a country commited to interacting in the competitative global economy. To this end, Oman has taken serious steps to develop higher education to cope with economic and market changes" (2007, p. 173). Altbach & Knight (2007) explain that in terms of knowledge transfer, signing this agreement entailed freely allowing cross-border provision of education (i.e., distance education and e-learning), education abroad (i.e., students travelling to other countries to study), commercial presence (i.e., opening educational facilities in other countries), presence of experts (i.e., qualified persons going to other

countries to exchange knowledge for a certain period of time). Since Omani higher education has not matured yet, its role in "knowledge transfer" could be described as being on the receptive side as a consumer of different forms of international education such as e-learning, foreign campuses, expatriate specialists, or sponsoring nationals to study abroad.

Reiterating the impact of globalisation on the Omani labour market, Donn and Al Manthri (2010, p.46) stress the importance of equipping gradutes with the skills needed to compete in the private labour market which is dominated by expatriates (e.g., Omanis constitute only 41% of the hotel industry). They report that "the government is very aware that each year more highly qualified graduates enter the labour market and that this must continue to occur if expatriates are to be replaced by suitably qualified Omanis". It has been pointed out that "there is an emphasis in the Sultanate on planning in line with labour market analysis at both the institutional and ministry levels" (Al Shemli, 2009, p. 16).

However, recent statistics show that the employment rate for the second cohort of Colleges of Applied Sciences graduates seven months after graduation stands at a worrying 10.54% of the total number of graduates in the six Colleges (see Table 1.2). The table shows that the International Business Administration (IBA) graduates are employed significantly more than their peers in Information Technology (IT), Communication Studies (CS) and Design.

Table 1.2. Employment Figures of CAS First Batch of Graduates until 31st of January 2011

Specialisation	**Graduates**	**Employed**	**Percentage**
Information Technology	521	57	10.94%
International Business Administration	215	43	20%
Communication Studies	308	16	5.19%
Design	236	19	8.05%
Total	1280	135	10.54%

Source: Career Guidance Department, personal communication, January, 2011.

Table 1.3. Number of Students in FP in the First Semester of 2010/2011 Categorised by College, Specialisation and Gender

College	Sohar		Nizwa		Sur		Salalah		Ibri		Rustaq		Total
Specialisation	M	F	M	F	M	F	M	F	M	F	M	F	
IT	98	98	-	-	147	40	-	-	32	143	17	31	606
IBA	-	-	-	-	-		192	22	-	-	19	200	433
Design	-	-	42	26	-	-	-	-	38	65	-	-	171
CS	-	-	87	182	89	49	69	25	-	-	-	-	501
English (education)	101	51	-	-	-	-	-	-	-	-	-	-	152
Total	199	149	129	208	236	89	261	47	70	208	36	231	1863

Source: CAS statistics of enrolled students at CAS FP in the first semester of 2010/2011, personal communication, 2011

The government states that higher education should be planned to closely match the needs of the labour market (Al-Lamki, 2006), yet the numbers in Tables 1.2 and 1.3 seem to suggest the opposite: the IBA graduates are almost twice as employable as the IT graduates, but CAS's intake into the IT specialisation was substantially more than its intake into IBA. Also, the CS graduates are the least employable, but the number of students admitted to CS in 2011 was the second highest. Such low employability rates in certain specialisations that have been deemed to be needed in the labour market have been partly attributed to the fact that these specialisations are offered by both public and private higher education providers, which has resulted in a premature saturation of these disciplines in the labour market; this had been expected by some authors (e.g., Donn & Al-Manthri, 2010), as mentioned earlier.

1.3.2. Globalisation and the English Language in Omani Higher Education

When discussing the role of the English language in the third world higher education sector, many attribute its wide and unprecedented spread to the colonial history of the countries in which it is spoken (i.e. the UK and the USA) (e.g., Alderson, 2009; Al-Issa, 2006), whereas others believe that globalisation has been the main driving force for its increasing use in higher education (Altbach & Knight, 2007; Donn & Al-Manthri, 2010; Pennycook, 1994, 1999; Phillipson, 1992). Al-Issa (2005), advocating the former view, states that transferring the language or values of a certain culture to another culture represents one aspect of colonisation, and argues that teaching English in Omani schools can be understood within this view. He explains that "one aspect of the national

curriculum is (first, second, etc.) language, which is a powerful tool for the transmission of the interests and values, concepts and beliefs of the dominant group" (p. 262). This view seeks to explain the current policies on the language of education in the context of colonial history.

Nonetheless, Al-Issa acknowledges the role of other factors such as the economic motives for mandating English as a second language in Omani education and the primary language in higher education. He stresses the important role that proficiency in English language plays in the process of 'Omanaising' the private sector (i.e., replacing expatriate labour with Omani labour). This view is more prevalent in the higher education literature, in which several authors argue that the proliferation of the academic higher education programmes taught in English worldwide can be associated with globalisation and attributed to economic incentives. Altbach and Knight (2007) maintain that the impacts of globalisation include the integration of research, the use of English as the lingua franca for scientific communication, the growing international labour market for scholars and scientists, the growth of communications firms and of multi-national and technology publishing, and the use of information technology (p. 291).

Attracting international students and promoting study programmes internationally have not, however, been the only motives for mandating the English language as the medium of instruction. In Omani higher education, a level of proficiency in English is a requirement to access most higher education institutions (HEIs), and English is considered a vital tool to access the national labour market (Al-Lamki, 1998, 2006; Donn &Al-Manthri, 2010; Al-Issa, 2006). Al Shemli (2009) looks at the role of English in higher education in the globalised context, and argues that "the main effect of globalisation in the Sultanate of Oman is the need to diversify the economy and raise standards; and the concomitant pressure to supply skilled graduates for rapidly changing economic conditions" (p.10). In this context, improving the English language skills of students is identified as a major challenge in higher education, though reforms have been undertaken at both the school and university levels (Al Shemli, 2009). Reforms in school

education that target improving proficiency in English language alongside skills in other subjects are highlighted by Alsarimi (2001), who calls for innovative methods to assess these skills and lessen the use of assessment tools that solely rely on memorising or rote learning.

> The new educational system aims to strengthen student competencies in mathematics and science, to improve student proficiency in English, and to teach students to use scientific methods and problem solving … to evaluate the richness of the diverse skills and knowledge in the new curriculum, it is crucial that student assessment be reformed as well (pp. 27-28).

As graduates' proficiency in English language is required by both the national and international labour markets, it has been identified as a vital asset in higher education. Though the internationality of English language as a lingua franca has also been emphasised as one of the reasons for this (Al-Issa, 2006; Al-Mahrooqi, 2012), the fact that the private labour market mainly operates in English has been seen by others as a more compelling reason. The need for graduates with an acceptable level of proficiency in English is clear in Al-Lamki's exploration of the barriers to Omanisation (i.e., replacing expatriates by Omani nationals in the labour market).

> Since English is the international language of communication and is also the medium for international business transaction, and since English is the operational language in Oman's private sector, it is recommended that the level and standard of English taught in schools and colleges be improved (2011, p.395).

In response, the governing bodies responsible for education in Oman have set conforming goals. The Ministry of Education states that:

> The government recognises that the facility in English is important in the new global economy. English is the most common language for international business and commerce and is the exclusive language in important sectors such as banking and aviation. The global language of Science and Technology is also English as are the rapidly expanding international computerised database and telecommunications networks which are becoming an increasingly important part of academic and business life

> (Reform and Development of General Education, Ministry of Education, 1995, p. A 5-1: as cited in Al-Issa, 2006).

The Ministry of Higher Education proclaim similar views on the role of English langauge in CAS. English language teaching is associated with national development in Oman; the National English language Policy/Plan (NELP) states that:

> the English language skills of Omani nationals must be seen as an important resource for the country's continued development. It is this recognition of the importance of **English as a resource for national development and the means of wider communication within the international community that** provides the rationale for English in the curriculum (Al-Issa, 2005 , p.2, emphasis in original).

1.3.3. Issues with Omani Students' Proficiency in English

Despite such stated intentions, plans and policies to promote the English language proficiency of the labour force, recent studies of graduates' English skills have found that these are inadequate for the needs of the private sector (Al-Mahrooqi, 2012; Al-Lamki, 2006). Al-Mahrooqi asserts that "research and experience have proved that the majority of school and college graduates possess neither adquate English language skills nor communication skills to function effectively in the workplace, which is dominated by expatriates from around the world" (2012, p. 124).

A similar view has been reported by the graduates themselves who "felt that their communication skills were very poor. Even the students on the verge of graduation expressed this, with much regret and sorrow" (Al-Mahrooqi, p.129). The students' conciousness of their lack of adquate language skills seems to have deterred them from applying for vacancies in the private sector; Al-Lamki reports that "students felt that the private sector discourages and disqualifies Omanis from applying because of the requirements for work experience and English language skills" (2006, p.392). She found that 72% of the 58 graduate students, in this study, considered English language a barrier to work in the private sector. However, it is suggested that one of the reasons for the low employability of nationals in the private sector had little to do with proficiency in English language or other skills, rather it was explained by Omanis reluctance to take

lower paid jobs (Donn and Al-Manthri, 2010). The issue of employability is very complex; one can only speculate that factors such as motivation, proficiency in English language or possessing other skills might be relevant, but the magnitude of these roles is still under-researched.

1.4. The English Language in the Colleges of Applied Sciences

In this section, the context of this study is described. The first sub-section describes how CAS, in which the study took place, was established and influenced by globalisation ideologies. The second sub-section explains the aims and components of the FP, and focuses on its assessment.

1.4.1. The Colleges of Applied Sciences

The CAS are state-sector colleges that provide free education to a limited number of secondary school graduates based on the students' academic merits and the colleges' capacity. Normally, there is an enormous demand for the places offered in the public sector HEIs in Oman, and being admitted to one of them has a great social value for the students and their families. There were originally six teacher training colleges spread across different regions, which were transformed into the CAS in 2005 to conform to the demands of the Omani labour market. The colleges maintained their separate locations to provide higher education services to a wider section of the population and local industries. In 2007, a royal decree was issued authorising the transformation of the former teacher training colleges into the present CAS. It proclaimed that:

> the transformation of the Colleges of Education at Nizwa, Sohar, Sur, Ibri, and Salalah into Colleges of Applied Sciences shall be approved commencing from the academic year 2005/2006. The Board of Trustee is permitted to transform the College of Education in Rustaq into a College of Applied Sciences or establish other new Colleges. Students who are subject to the Basic Law of Colleges of Education shall continue thereunder until they graduate (CAS, 2010a, p. 3)

CAS is governed by the Ministry of Higher Education, which places Omanisation and preparing employable graduates as two of CAS's priorities.

> Omani government has set targets for Omanisation in line with the wishes of His Majesty Sultan Qaboos bin Said who has said that 'Omani youth constitutes a large vital section of the society and no effort should be spared to ensure a bright, dignified future for them'. This bright future can be achieved by gaining a degree from CAS and finding work in the field of your choice. (CAS, 2010e, p.2)

The study was conducted in two Colleges: Sur and Rustaq. CAS in Sur is located about 200Km in a coastal town while CAS in Rustaq is located 120 Kms from the capital city in an interior town. The colleges offer different academic programmes, Rustaq College offers Information Technology, International Business Administration, and English Language (education) Programmes, while Sur College, offers Communication Studies and Information Technology programmes. However, the shared programme (IT) uses similar books, curriculum and assessment instruments. The gender distribution of the students in the two colleges is not equal. There are more male students in Sur College than female students while there are more female students in Rustaq College than male students. The difference in gender distribution between the two colleges can be simply explained by the capacity of these colleges to provide in-campus accommodation for female students. The scholarships provided by CAS include providing in-campus accommodation for female students and housing allowance for male students. Rustaq College has a larger capacity to provide accommodation for female students; therefore, the number of female students is higher (see Table 1.3 above).

Taking CAS as an example, Donn and Al-Manthri (2010) discuss the growing trend of steering tertiary education towards building human capital and supporting the economic goals of countries as is the case in Oman and the other Arab Gulf States. To achieve this, these countries usually rely on foreign experience in higher education. This process, the authors argue, is a form of 'policy borrowing', which means bringing in curricula, teaching methodologies, ideologies, assessment methods and other policies from foreign

institutions which are internationally recognised as being advanced in the field of tertiary education. Policy borrowing is one of the dissemination methods through which globalisation driven ideologies are spread (Ball, 1998; Dale, 1999).

In CAS, the English language was chosen to be the language of instruction when various English speaking higher education 'policy entrepreneurs', as Ball (1998) calls them, were invited to put forward their proposals and plans for the six amalgamated Colleges. In 2006, the Ministry of Higher Education, under which the Colleges operated, signed a contract with Polytechnics International New Zealand (PINZ) to conduct a needs analysis of the labour market and recommend the future academic programmes of the colleges. The programmes offered by the colleges currently, as a result of the PINZ report, are IT, Design, IBA and CS. This approach to creating new HEIs has been criticised for being totally foreign to the local cultures; Donn and Al-Manthri argue that "they [the Gulf countries] have little control, other than as purchaser and consumer, over the language or the artefacts of the language" (2010, p.24). When the programmes the colleges would offer were agreed upon, New Zealand Tertiary Education Consortium NZTEC was contracted to provide the curriculum as well as part of the assessment and other services. The first batch of the students had to go through an English language preparation programme (i.e., FP) for almost an academic year before qualifying to take the academic courses in English.

1.4.2. The Foundation Programme

It is estimated that almost 80% of the students admitted to higher education in Oman require English language courses in the FP before starting their academic study (Al-Lamki,1998). The FP is a pre-sessional programme that can be considered an integral part of almost all of the HEIs in Oman. Its general aim is to provide students with the English language proficiency, study skills, computer and numeracy skills required for university academic study (OAAA, 2009). The aim of teaching English language is stated to be "equip[ing] students with both the English Language and academic study skills they will need to succeed in their subject studies" (CAS, 2010e, p.33). The FP

consists of twenty hours per week of English language instruction, and two hours of mathematics and/or computer skills courses in each semester. The English language programme is divided into two major courses, the General English language (GES) and Academic English Skills (AES) as shown in Table 1.4. In this study, the term (FP) will be used to refer to the English language components only of the programme.

Table 1.4. English Language Courses in the Foundation Programme and their Approximate Equivalent Levels in IELTS

Equivalent in IELTS	**Foundation Programme Levels**	**Courses**	**Weekly Contact Hours (Hrs.)**	**Total Hrs.**
IELTS 3.0 or below	Level C	GES	11	20
		AES	9	
IELTS 3.5	Level B	GES	11	20
		AES	9	
IELTS 4.0	Level A	GES	11	20
		AES	9	
IELTS 4.5	Entry to First Year	EAP	10	10

Source: modified from *Colleges of Applied Sciences Prospectus*, (2010, p. 33)

1.4.3. Language Assessment in the Foundation Programme

A common concern that is raised about pre-sessional programmes in general is that they allow students to embark on academic study with an inadequate level of English proficiency (Allwright & Banerjee, 1997; Fox, 2004). Cotton and Conrow (1998) report that, in their study, students expressed a need for extra EAP instruction even after they had reached the IELTS level required by their universities. Though most internationally recognised higher education institutions do not permit embarking on higher education before reaching a certain minimum level of English language proficiency, some others do allow students to start academic studies at lower levels of language proficiency and provide them with language support programmes (Fox, 2004). CAS follows this approach: students are provided with two EAP courses in their first year and two English for Specific Purposes courses in their second (one each semester) to help them overcome some language challenges they might face when starting academic study.

The present assessment system in CAS uses both standardised tests and Continuous Assessment (CA) as the way forward for education and assessment reforms in Oman. Alsarimi (2001, p.28) gave a rationale for this and argued that the reform in education should be matched by a reform in assessment instruments, and placed special emphasis on using CA to assess students' skills:

> The new Omani educational system, on the other hand, advocates diversified assessment techniques, with more emphasis on authentic student assessment. When implementing the new educational system, teachers are expected to: (a) put less emphasis on simple memorisation of content and final paper-and-pencil examinations, (b) teach by applying knowledge and materials to the lives of the students, include higher-order-thinking, (d) and use CA or on-going assessment methods.

In support of this argument, an empirical study of the relationship between the type of assessment and academic achievement in the Omani school system context (Al Kharusi, 2008) suggested that 'alternative assessment' results in better achievement than traditional tests; especially when it is used by experienced teachers.

> Armed with the classroom assessment literature regarding the advantages of alternative assessments as well as with the achievement goal research regarding the potential negative consequences of adopting performance-approach goals, the present study findings tend to support the movement towards the use of more alternative assessments (p.262).

Al Kharusi (2008) defined alternative assessment as "another title used for describing performance assessments to indicate that they are alternative to traditional assessments" (p. 245). The recommendation to integrate CA as part of the overall assessment was considered, adopted and enforced by CAS academic regulations. In a description of the assessment used in CAS assessment, it is stated that "academic regulations mandate the allocation of 50% of marks to a final examination and 50% to CA" (*CAS*, 2010e, p. 35).

In the FP, students take two courses in which they undergo two different assessment instruments. Table 1.5 shows that assessment in the GES course includes a mid-term test and a final test, whereas assessment in the AES course includes writing a report and

presenting it orally. In order to pass, students must obtain 50% of the total marks in each course; failing to achieve this means failing the FP and consequently being denied access to higher education.

Table 1.5. Assessment Instruments in the Foundation Programme Courses

Course	Assessment Instruments	% of Course Total	% of Foundation Programme Total
General English Skills	Mid-term Test	40%	50%
	Final Test	60%	
Academic English Skills	Presentation	50%	50%
	Report	50%	

In CAS, English language assessment is a product of multiple factors such as globalisation, educational trends and contextual issues all of which are discussed in the next section.

1.5.Rationale of the Study

English language assessment in Oman has not only been influenced by the approaches and techniques in the field of educational assessment, but also by international trends. Therefore, in investigating the effectiveness and predictive validity of the FP English language assessment, policies on higher education language assessment should all be considered along with pertinent theoretical and empirical literature on language assessment. In doing so, this study hopes to contribute not only to understanding the effectiveness of FP assessment in its local context, but to add to the wider knowledge on the influences of globalisation on international higher education policies.

Passing the English language component of the FP assessment gives access to public higher education after meeting other academic requirements. With the increased stakes of assessment, the issue of validity becomes more important and interpretation of the assessment scores is central to this validity (Messick, 1989). In the Foundation Programme at CAS, both continuous/performance assessment and tests are used to measure students' language skills; though this combination has been advocated by some authors arguing that it increases assessment validity and results in better academic achievement (e.g., Alsarimi, 2000; Al Kharusi, 2008; Hamilton, 2003), it has been

criticised by others who believe that the incompatible views of validity advocated by each type could result in distorting the scores' interpretation (e.g., Teasdale & Leung, 2000). Another area that provokes varying responses is the association between language proficiency and academic achievement or what is known as the predictive validity of language assessment. Previous research on this topic has reported conflicting findings to the extent that some have claimed that this is not a fruitful line of research (see Section 3.3). All of these arguments are central to language assessment in higher education and were considered in exploring the effectiveness and predictive validity of language assessment and how students and teachers perceive them.

1.6.Questions of the Study

To investigate all of the areas mentioned above, this study was designed to be conducted in two phases. In the first phase, questionnaires (of two kinds), interviews and focus groups were used; the students on the FP were asked to complete a questionnaire and participate in focus groups; the teachers were asked to complete a different questionnaire and to take part in an interview. In addition, textbooks, test papers, assessment activities and policy documents were analysed. In the second phase, the students who had now started the first academic year were asked to participate again by responding to a questionnaire and participating in focus groups. The teachers of the academic and English language courses that the students took were interviewed and asked to respond to a questionnaire as well. In this second phase, courses syllabi, test papers, textbooks and students' scores in the English language and academic courses were also analysed.

Designing the study to be conducted over two semesters was necessary to capture and understand the experiences the students went through when they moved from studying language courses only to studying academic courses. It has also been suggested that the first semester of an academic study yields more information about the predictive validity of language assessment than do the following semesters (Graham, 1987); the power of the predictive validity in the first semester of an academic study is usually higher than it is in the subsequent studies:

> An effective test of language competence …would be one based on the language demands of the first six months of Tertiary education (Phillips, 1987, p. 78).

The questions that this study investigates are listed below.

1. How well did the process of assessing students' English language performance, through CA and tests, function[5] in the Foundation Programme?

   1.1. What processes and procedures followed in writing and implementing the assessment instruments as depicted by the official documents?

   1.2. How was the reliability and validity of FP assessment viewed by students and teachers?

   1.3. How was the impact of FP assessment perceived by students and teachers?

   1.4. What were the differences between the 'CA' model used in the Academic English Skills course and the 'test' model used in the General English Skills course in terms of effectiveness, accuracy, and preferences of students and teachers?

   1.5. How did teachers perceive the centrally controlled assessment used in CAS?

   1.6. What types (criterion/norm-referencing) of assessment were used? And how?

   1.7. In all the above, were there any significant differences between the views of the students' groupings by college, gender, age, self-evaluation and teachers' groupings by college, gender, college, age, nationality, teaching and assessment experiences?

2. How did the assessment instruments correspond to the stake-holder wishes?

   2.1. What were the national and international policies on teaching and assessing language that influence assessment in Oman? And how does FP assessment correspond to these policies?

   2.2. What were the student and teacher perceptions of the assessment tools' effectiveness and their roles in shaping language assessment in retrospect?

[5] The verb function indicates how FP assessment was (should be) designed, implemented, used, and viewed.

3. What was the predictive validity of the English language assessment for student performance on the academic courses?

   3.1. Did student performance in English language assessment on the FP correlate positively with their performance in academic courses?

   3.2. Did the strength of correlation between the language proficiency and academic achievement differ significantly when students' scores in English language tests only or CA only were used, instead of the overall scores in both?

   3.3. Did the groupings by college, gender, self-evaluation and specialisations show significant differences in the correlations between language proficiency and academic achievement?

   3.4. How demanding were the learning outcomes and assessment of the academic courses in the FY on students' language skills?

4. How did the stakeholders understand the relationship between the student performances in the English language assessment and their performances in the academic courses' assessment?

   4.1. What were student and teacher perceptions of issues related to the design, marking and impact of the English language assessment?

   4.2. How did students and teachers think language accuracy should be considered in assessing academic assignments?

   4.3. What were student and teacher perceptions of the importance of the predictive validity?

1.7.Thesis Outline

This thesis consists of 11 chapters. Chapter 1 has presented the role of English language in higher education and discussed the factors that have influenced policies in higher education, including globalisation. Then it has provided background information about higher education in Oman and related policies on English language education and use as the medium of instruction in most higher education institutions. The FP and its assessment at CAS were then discussed as background to the study.

Chapters 2 and 3 review literature on the topics investigated by the study. Chapter 2 explores the literature on language testing and assessment, particularly classification of tests and the differences between tests and assessment. The first section describes test purposes, types and approaches. The second section of the chapter examines the epistemological and validity considerations inherent in the premises of testing and assessment. Chapter 3 discusses and links three wide areas: language programme evaluation; language assessment validation; and predictive validity of language assessment. The first section covers arguments about, and views on, programme evaluation including definitions, types, approaches, and epistemological paradigms in the field. The second section explores conceptualisations of language assessment validity and validation, focusing on some proposed models for undertaking validation studies; it also looks at the interconnectedness between the fields of programme evaluation and assessment validation. The third section focuses on empirical studies conducted on the predictive validity of language assessment, and it surveys the findings reported in some previous studies on the predictive validity of English language assessment instruments including IELTS, TOEFL and in-house tests, and discusses possible factors that might explain the variations in reported results and the methodological limitations of previous research studies.

Chapter 4 presents the design of this study and justifies using certain methods by linking them to the study questions and purposes. It discusses the epistemological basis of this study and pragmatic arguments for using a mixed-methods design. It provides background information on the study's participants, location and phases. It also describes the stages that the researcher went through in designing and piloting instruments and procedures, and collecting and analysing data. The last section of the chapter discusses issues of study quality, ethical considerations and limitations.

Chapter 5 presents the results obtained by analysing documents related to FP assessment. It provides an in-depth analysis of a number of different documents ranging

from textbooks, course syllabi, test papers, test instructions and handbooks to policy documents. Using thematic analysis, the results are categorised into four central topics (1) norm versus criterion-referenced assessment, (2) incompatibility between what is assessed and what is taught, (3) inconsistency in implementing assessment criteria, and (4) replication of general foundation programme standards standards in FP specifications. These findings are discussed briefly in the last section of the chapter and links are made where appropriate to previous comparable studies.

In Chapter 6, the results obtained from the first phase of student and teacher questionnaires are presented and discussed. These questionnaires are used to provide an overview of student and teacher perceptions of the assessment instruments used in the FP. The results of the questionnaires are categorised into four areas: (1) student and teacher views on assessment validity, reliability and satisfaction, (2) assessment impact, (3) tests versus CA, and (4) centrality of assessment writing. Significant differences between the groups are identified and investigated for possible implications. Results from both the student and teacher questionnaires are compared and discussed in the last section of the chapter.

In Chapter 7, student and teacher views expressed in the focus groups or interviews in Phase 1 are analysed and the results are categorised into common themes. The students' and teachers' perceptions shared five main themes: (1) uncertainty about assessment details, (2) perceptions of assessment effictiveness, (3) perceived need for more assessment instruments, (4) comparison of tests to CA, and (5) assessment impact.

Like Chapters 6 and 7, Chapters 8 and 9 present the findings from student and teacher questionnaires, focus groups and interviews which were conducted in Phase 2 of the study. Though most of the students who had participated in Phase 1 also participated in Phase 2, the participant teachers in the two phases were different. In Phase 2, teachers from the academic departments as well as the English language department were invited to participate, while in Phase 1, only teachers from the English language department

were invited. The results from the instruments used with the students and teachers are presented, analysed and compared. The differences and similarities between the students' and teachers' perceptions are discussed in the last section of each chapter with some references made to related literature.

Chapter 10 presents the results of FP assessment predictive validity. A correlation study was conducted to investigate the relationship between students' scores on the FP and their scores in the FY. In this chapter, the differences in the strength of the predictive validity across specialisation and self-evaluation groups are focused upon and some possible explanations suggested.

Chapter 11 has two main sections. The first is a discussion of the findings reported in the previous chapters. It discusses the findings on two main points raised by the study questions, the effectiveness and predictive validity of the FP assessment. Study findings on the effectiveness the FP assessment are evaluated using the validity theory proposed by Messick (1989), particularly to explore any signs of threats to validity: construct irrelevance or construct underrepresentation. An argument about the effectiveness of FP assessment that incorporates evidence from different sources is then introduced; the findings on the FP assessment predictive validity and its effectiveness constitute the evidential and consequential basis of this argument. The chapter also revisits and attempts to explain other interesting findings presented in previous chapters and link them to relevant literature.

In the second section of this chapter, the implications of this study are discussed. These implications are divided into practical, theoretical and policy related implications. The last section of this chapter considers the limitations of this study and provides some suggestions for future research.

1.8.Chapter Summary and Conclusion

This introductory chapter has outlined the demographic context of this study. It started with a narration of an incident in which the power of assessment provoked student demonstrations against elements of assessment systems at higher education institutes.

From this starting point, the chapter discussed how English language education in higher education is one aspect of globalisation. It argued that different countries have different motivations for increasingly making their higher education an English language medium education. It then presented the case of Omani higher education in which the English language is taught as a graduate asset for future employment, but recent evaluations of students' English language abilities have shown that more effort should be put into equipping students with the language skills required in the labour market, and it is also widely felt that language assessment should be reformed as should language teaching.

The following chapters will provide an exploration of the literature on language testing and assessment. It will review studies on English language assessment validation, evaluation and predictive validity.

Chapter 2: Language Testing and Assessment

2.1. Introduction

This chapter provides a review of some of the literature on language testing and assessment and focuses on concepts and issues most relevant to the empirical work in the study. The chapter is divided into two main sections. The first (Section 2.2) presents current classifications of tests [this word except where otherwise stated, will be used to cover all assessment instruments, but a narrower meaning will be explored later in the chapter in terms of their purposes, types and approaches]. These classifications refer to tests; however, they are applicable to both tests and other assessment instruments. The second (Section 2.3) examines the differences between standardised tests and performance assessment as measuring tools of students' language proficiency. These differences are linked to epistemological and validity considerations, then some arguments about combining marks obtained from standardised tests and other types of assessment are explored in the last section of this chapter.

2.2. Classifications of Tests

There are a number of ways in which assessment instruments can be classified. They can be categorised according to their purposes, their construction, and their use.

2.2.1. Test Purposes

Over the last three decades, classification of the purposes of language tests into four main types has generally been followed, namely: placement, diagnostic, achievement, and proficiency (Harrison, 1983). Additions or changes to this classification were suggested by more recent authors, notably Bachman (1990) added "entrance tests" as a fifth type. Also aptitude tests, which were popular in the past and used to predict the likelihood of students' success in learning foreign languages, are not mentioned in most current textbooks on language testing (e.g., Hughes, 2003). In this section, these five types are discussed and different views on their meanings or uses are explored.

2.2.1.1. Proficiency Tests

Proficiency language tests are designed to measure candidates' ability in a language in general without special regard to any prior courses taken or training undergone (Hughes, 2003). Hughes says that this type of test aims to identify whether a candidate has reached a specific level of mastery or not. Similarly, Weir (2005) states that a proficiency test is "a test which will provide information on a candidate's ability to perform in a future specified target situation". Harrison (1983) explains that proficiency tests are concerned with "a student's ability to apply in actual situations what he has learnt" (p. 7), but he stresses these tests are not necessarily based on a course and are usually are concerned with future needs; he argues that this test is the best type for admission purposes (e.g., a university language entrance test).

2.2.1.2. Achievement Tests

This type of test measures a candidate's ability to reach the objectives set for a specific course. Hughes (2003) identifies two sub-types, the *final* achievement test and *progress* achievement test; both of these tests are based on specific course objectives. The first type is administered at the end of a course while the second is administered during the course to measure progress in achieving some objectives. The other difference between the two is that a *final* achievement test is a summative test while a *progress* achievement test is a formative test (Hughes, 2003). This definition seems to agree with that of Harrison (1983) who labels the final achievement test an "attainment or summative" test.

One contentious topic in achievement testing is the content of an achievement test. Should this test be based on course objectives or taught materials? Hughes says:

> tests based on objectives work against the perpetuation of poor teaching practice ... It is my belief that to base test content on course objectives is much to be preferred; it will provide more accurate information about individual and group achievement, and it is likely to promote a more beneficial backwash effect on teaching (2003, p. 14).

He acknowledges that this might be unfair for students using unsuitable course books, but is better for the long term improvement of the course. However, this view has been criticised by Weir for disregarding the teaching that occurs in the classroom and relying on positive "washback" to amend poor teaching or a poor curriculum. In this regard, Weir argues that

> classroom testing should not be divorced from the teaching that precedes it. Achievement testing should be firmly rooted in previous classroom experiences in terms of activities practiced, language used, and criteria of assessment employed … The purpose of tests of achievement should be to indicate how successful the learning experiences had been for the students rather than to show in what respects they were deficient (1993, p.5).

He adds that "we must also ensure that the students are adequately prepared for the tasks they will have to face" (1993, p. 6), and warns that if students are faced with tasks that they have not practiced before, they are more likely to underperform. Like Weir, Bachman advocates using a syllabus to guide the content of an achievement test, stating that

> while the specific types of tests used for making decisions regarding progress and grades may vary greatly from program to program, it is obvious that the content of such tests should be based on the syllabus rather than the theory of language proficiency. That is, they will all be achievement tests (1990, p. 61).

To summarise, the view that an achievement test should be based on course objectives without considering the teaching materials or learning that occurs in the classroom seems to be more concerned with creating positive washback than assessing students' language proficiency. More pervasive, however, is the view that we should base an achievement test on taught materials and ensure that teaching materials correspond with course objectives to be fairer to students and achieve better results.

2.2.1.3. Diagnostic Tests

This type of test aims to identify candidates' strengths and weaknesses in a certain area and inform decisions on whether more instruction is needed. Linn and Miller (2005, p.41) limit the use of this type to identify the "causes of persistent learning difficulties", whereas Hughes (2003) argues that proficiency tests can sometimes be used as diagnostic tests, but stressed that the latter should be more detailed and reliable, and should provide a wide coverage of the abilities tested. Harrison (1983) describes the function of a diagnostic test as a formative or progress test, which provides information on the progress of students and the remedial work to be undertaken. Harrison also differentiates achievement tests from diagnostic tests in that the former look into a longer period of time that can sometimes include one or more courses. Bachman (1990) says though all tests have a diagnostic element in finding the weaknesses and strengths in a candidate's proficiency, diagnostic tests are "developed specifically to provide detailed information about the specific content domains that are covered in a given programme or that are part of a general theory of language proficiency" (p. 60). The specificity and amount of details of this type of test is partly why they are scarce. Hughes (2003) identifies DIALANG as one of a very few international standardised diagnostic tests. DIALANG is a "diagnosis system" that informs on student levels against the Common European Framework for language learning. It includes items on five categories or skills: reading; writing; listening; grammar; and vocabulary. In general, however, diagnostic tests are often developed locally to meet instructional purposes.

2.2.1.4. Placement Tests

Placement tests are used to identify different groups of candidates for future placements and their content is based on future courses. Hughes argues that the best placement tests are those used for the specific purposes for which they were designed (2003). Unlike Hughes, Harrison (1983) argues that placement tests address current general abilities and are not related to future courses. Bachman (1990) brings together the two opinions stating that specifications for a placement test can be drawn either from a proficiency

theory or the learning objectives of a syllabus to be taken. He argues that the number of students to be admitted to a programme should be the criterion for selecting the content of a placement test. To maintain steady quantities of students entering a certain programme, he advises using norm-referenced placement tests, but if a programme's intake capacity is flexible then criterion-referenced placement testing is a better option.

2.2.1.5. Test Tasks and Test Purposes

The following table, from Harrison (1983, p. v) lists a number of test tasks and how useful they might be for particular test purposes as proposed.

Table 2.1. Usefulness of Test Tasks for Specific Purposes

Test Type	Placement	Diagnostic	Achievement	Proficiency
Scripted speech + true false items	1	3	3	3
Narrative text + true/false items	1	3	3	3
Structured writing	1	2	2	2
Cloze	1	x	2	2
Dictation	1	2	2	2
Conversation	1	x	2	2
Scripted speech + multiple choice pictures	x	1	3	x
Scripted speech + completion items	x	1	3	x
Completion + writing	x	1	2	x
Completions + multiple choice fillers	x	1	3	x
Transposition	x	1	2	x
Unscripted speech + multiple choice items	2	3	1	2
Unscripted speech + visuals	2	3	1	1
Text and argument + multiple choice items	2	3	1	2
Letter	2	3	1	2
Re-orientation	x	2	1	x
Speaking to pictures	2	2	1	3
Talking on a topic	2	x	1	1
Transfer	3	3	2	1
Following instructions	2	2	2	1
Giving advice	x	2	3	1
Appropriate response	x	3	2	1
Sequence	x	3	3	1
Role play	x	2	2	1
Problem solving	x	x	2	1

Note: The numbers indicate how useful each type of test is likely to be for the four purposes, placement, diagnostic, achievement and proficiency, ranging from 1 (most useful) to 3 (useful only in some circumstances); x means not suitable for this purpose.

These test tasks can be considered as a starting point from which different types of tests could be constructed. These are particularly useful for constructing tests in the Omani context where the Foundation Programme includes two types of tests: a placement test administered at the beginning of the academic year and an achievement test administered at the end of each semester. The previously presented discussion of test classification assists in understanding and analysing the tests, test tasks, test writing and administering procedures used in FP assessment. However, it should be noted here that the classification of test tasks presented by Harrison (1983) is for guidance and might not always be feasible or applicable in actual language tests. For example, "giving advice" can be used as a test task in a placement test unlike the suggestion that it is not suitable for this type of tests.

2.2.2. Approaches to Test Construction

In language assessment handbooks, approaches to test construction are another way to categorise tests (e.g., Harrison, 1983; Hughes, 2003; Weir, 1993). This section will discuss the differences between direct/indirect and objective/subjective testing. It will then explore issues concerning the use of rating scales and focus on one extensively researched topic: raters' variability.

2.2.2.1. Direct and Indirect Testing

The term "direct testing" is sometimes used when candidates are asked to perform the skill directly using tasks such as writing or speaking. Hughes (2003) claims that this type of testing is better for proficiency and achievement purposes in which decisions need to be made about students' mastery levels and abilities, provided that a wider selection of the tasks is used to give a more accurate indication of assessed abilities.

The indirect approach to testing uses proxy measures of language abilities; one example of such tasks using the indirect approach is asking students to fill in a gap after reading a script to measure their reading skill. Hughes states that "indirect testing attempts to measure the abilities that underlie the skills in which we are interested" (2003, p.18).

Arguably, scores in indirect testing can be more general than direct testing: the latter focuses on a set of tasks that might not be representative of all constructs evaluated. However, the relationship between the indirect test tasks and the abilities tested tends to be weak (Hughes, 2003). Weir (1993) argues that it is more useful to assess students using direct approaches than indirect ones as the former assist students to obtain skills needed in future studies, such as writing or presenting.

2.2.2.2. Subjective versus Objective Testing

Objective testing, in its usual meaning, entails scoring that does not need any significant judgement by the scorer. Harrison (2003) says that it involves tasks with only one correct answer, whereas subjective testing uses scoring systems that rely on the scorer's judgement where more than one correct answer is possible. Of course the first type is more reliable, but the second type's reliability can be increased by using well-trialled marking scales, and by training scorers. Harrison (1983) urges trialling the marking scales before use to allow for any necessary amendments.

Subjectivity and inconsistency in scoring written scripts is the most extensively researched topic on subjective testing. Bachman (1990) points out however, that subjectivity is not limited to scoring procedures but is also evident in every procedure in the test and in the actions of all test users. For example, test developers make subjective decisions about what tasks to use and how to order them and students make subjective decisions about, for example, how to undertake the tasks.

2.2.2.3. Using Rating Scales

Developing appropriate rating scales is important for creating effective assessment instruments because of the feedback they provide to students. A rating scale is "a set of guidelines for the application of performance criteria to the responses and performance of the students" (Linn & Miller, 2005, p. 261). One classification of rating scales is Hamp-Lyons (1991), in which they are divided into three types: holistic scales; primary

trait scales; and multiple trait scoring. This classification was followed by Fulcher (2010, p.208) who explains that a holistic scale looks at "overall quality of the performance", a primary trait scale awards a single score which "reflects the specific qualities expected in writing samples at a number of levels on the scale", and a multiple trait scale "requires raters to award two or more scores for different features or traits of the speech or writing sample". Like several writers in the field of language assessment, Fulcher stresses the need to ensure inter-rater reliability and intra-rater reliability; this can be achieved by implementing moderation and standardisation policies. Weir (1993) explains that the moderation stage of a test includes reviewing test tasks in terms of the level of difficulty, discrimination, appropriateness of sample, overlap, clarity, timing, layout, and examination of bias and the procedure also includes specifying the marking criteria for each task. Standardisation aims to "bring examiners into line, so candidates' marks are affected as little as possible by the particular examiner who assesses them" (Weir, 1993, p. 28). Moderation and standardisation aim to achieve higher levels of reliability and lessen rater inconsistency which is the topic of the next section.

2.2.2.4. Rater's Invalidity/Variability/Inconsistency

An aspect of subjective testing that has been extensively studied in language assessment is rater's inconsistency in using and interpreting marking scales. Clapham (2000) criticises the low reliability of some marking scales resulting from them not being trialled before use, and argues that in such cases the tasks and marking schemes become invalid. Banjeree and Wall (2006) report variability in the way teachers understood rating scales in their study, and state that "the interviews revealed that the route by which the tutors arrived at their summary judgments differed" (p. 63). One of the recommendations of this study is addressing and minimising this variability by using training sessions prior to actual marking.

Several other studies have reported similar findings about rater's inconsistencies in different contexts and four examples will be considered here. In Eckes (2008), 65 expert raters in German as a foreign language were asked to rank nine routinely used

descriptors in terms of importance. They ranked the nine criteria significantly differently in terms of "general importance" and the "importance for scoring examinee performance". The author claims that this finding has confirmed previous findings on raters' variability.

Elder, Barkhuizen, Knoch and Randow (2007) attempted to improve rater reliability in marking the writing component of a diagnostic assessment used to determine the needs of undergraduate students in a New Zealand University. The researchers used immediate online feedback in the form of a discrepancy score showing the difference between the raters' score and the 'official ratings'. Eight raters were asked to mark a number of scripts which had already been rated, benchmarked and moderated. They marked the scripts twice, once before and once after receiving the feedback. The results revealed "limited overall gains in reliability", but "there was considerable individual variations in receptiveness to the training input" (Elder et al., 2007, p. 37). Despite the limited improvement in the raters' reliability, the authors regarded the approach implemented as "promising" in enhancing reliability levels among raters.

Lumley (2002) investigated the strategies teachers followed while using a marking scale with four different categories to mark two writing tasks in a test that intended to assist the Australian government in immigration decisions. He reported that the scales were sometimes used to justify teachers' decisions, convey their judgments or as an instrument to "narrow" their evaluation of a written piece. He argues that when teachers are faced with a complicated situation they tend to resort to other ways to mark a script such as comparing the scripts or placing more emphasis on a certain criterion. He concluded that "the scoring decision appears not to be based on the scale. Such behaviours recur - in disparate and unpredictable ways - with all four tasks examined in this study, and with all four raters" (2002, p. 262).

To understand the differences between using a holistic scale and a detailed one, Knoch (2009) recruited ten raters and asked them to mark 100 scripts that had been produced by

a large-scale diagnostic test for native and non-native speakers in a University in New Zealand. The raters were asked to use a holistic scale and a detailed one. The raters then filled in a questionnaire, after which they were interviewed about their perceptions of the functionality of each of the scales. The results revealed that the rater reliability was better when the detailed scale was used, and that the holistic type rating often resulted in a halo effect, where "a rater awards the same score for a number of categories on the scale" (2009, p. 293). It was found that the *halo effect* also occurred when encountering difficulties in rating.

One of the factors that have been widely investigated in rater variability is their backgrounds. Johnson and Lim (2009) studied the differences between native speakers and non-native speakers' rater reliability. The study focused on a test administered to speakers of English as a second/foreign language and used to assist in university admission decisions in the USA. The sample included 19 teachers, four of whom were non-natives. The results showed minimal differences between all teachers, and no language group specific differences were found. They argue that "ratings in this performance assessment of writing are on the whole accurate, reliable, and fair" (2009, p. 500). On the same topic, Brown (1995) explored the possible effect that raters' backgrounds could have on rater variability. In this study, 51 raters were recruited and were grouped according to their experiences in tour guiding and/or teaching into three groups: guiding experience only, teaching experience only, and both guiding and teaching experiences. They were asked to rate a number of speaking activities that constituted multiple phases. The results revealed that there were differences among the raters in rating the activities, but that these differences were 'minor' and non-significant.

The importance of identifying these approaches of test construction is apparent in the fact that FP assessment uses a mixture of approaches such as direct, indirect, subjective and objective testing. This presentation of what these approaches involve and how they influence tests and test-tasks will facilitate a later discussion of the findings on FP assessment.

2.2.3. Test Types

After reviewing some of the literature on test classifications based on their purposes and approaches, this section presents different test types based on their uses such as formative/summative tests and criterion/norm-referenced tests. It also discusses the meaning of outcomes-based assessment and highlights its ties with politically driven uses of language tests.

2.2.3.1. Formative and Summative Assessment

Some authors distinguish between these two types, indicating that summative assessment occurs at the end of a course to meet institutional requirements (e.g. measuring achievement), whereas formative assessment is used to guide students in the learning process or assist teachers adjust their lessons, teaching materials or methods to best suit the needs of the students (e.g. Bachman & Palmer, 1996; Brindley, 1998; Yorke, 2003). Yorke states that "the central purpose of formative assessment is to contribute to students learning through the provision of information about performance" (2003, p. 478). However, he seems to believe that, in the UK context, this purpose has been understated and underused because of the "unitization of assessment" and implementation of more summative assessment instead. This, he argues, results in providing "insufficient" or "late" feedback. He points out that the effectiveness of formative assessment in enhancing the learning experience is dependent on the quality of the feedback given to the students (Yorke, 2003). Though it is sometimes assumed that performance assessment typically serves formative purposes while tests serve summative purposes, actually each could serve either purpose. Nitko (1995) argues that performance assessments are similar to tests in that they can be constructed to serve formative or summative purposes and stresses that the curriculum should be the ultimate basis for setting performance assessment activities and designing marking criteria, as it should be for tests too.

Brindley (1998) claims that formative assessment, although surely needed, has largely disappeared in many contexts and this is an unfortunate development. He maintains that "the political reality seems to be that when there are two competing assessment schemes,

system information needs will override those of formative assessment" (p. 61). He also claims that the political drives for summative assessment and the pedagogical need for formative assessment generate a dilemma. He observes that "a number of commentators have recently highlighted the inherent dilemma in trying to reconcile demands for national comparability with the need to relate assessment directly to the learning process" (1998, p.47). The differences between summative and formative assessment emerge again in the following two sections which discuss criterion/norm-referenced assessment and outcome-based assessment.

2.2.3.2. Distinction between Norm-referenced and Criterion-Referenced Assessment

Generally assessment instruments are used for either norm-referenced, or criterion-referenced purposes depending on stake-holders' or institutions' needs. Norm-referenced testing (NRT) "relates one candidate's performance to that of the other candidates. We are not told directly what the student is capable of doing in the language" (Hughes, 2003, p. 20). Criterion-referenced tests (CRT) aim to "classify people according [to] whether or not they are able to perform some task or set of tasks satisfactorily" (Hughes, 2003, p.21). Davies (1990) recognises norm-referenced use of tests, but stresses that it includes an element of *imposing a normal distribution*. He argues that NRT "imposes a normal distribution on those under test, whether or not such a distribution is there in reality" (p.17). This suggests that NRTs are inadequate in situations where students' scores are not expected or intended to form a normal distribution in which 50% of the students fall in the middle range of the scores. Other writers such as Brown (1990) stress the value of NRT in measuring students' proficiency in general language skills. Also, Hughes (1986) underlines the issue that the norm-referencing procedures specifically used in item analysis or for measuring reliability are "well established" in many educational contexts. Fulcher (2010) takes this argument further and claims that the "paradigm of *norm-referenced* testing is the normative approach in educational testing generally" (emphasis in origin, p.31). In school education, Gipps (1999, p.289) claims that using norm-referencing assessment on national or international levels could produce negative outcomes; she states that "the pre-occupation that there is in so many countries

with comparison … is for most children … educationally inappropriate and in many cases damaging, as children learn to lower their self-esteem and switch off from achieving" (p. 289).

On the other hand, CRTs measure the test takers' performance against a set of attainable goals or outcomes which are determined in advance (Alderson et al., 1995; Davies, 1990). Fulcher (2010) places the first discussions about CRTs in the 1960s and claims that CRTs are more useful than norm-referenced tests when the required information concerns pedagogical issues. Similarly, Brindley (1989) argues for using CRTs to evaluate students' achievement in a specific course. However, Davies (1990, p.19) claims that "a criterion-referenced test is one use of a norm-referenced test" and that the two are not completely different. Unlike Davies, Brown (1990) argues that tests' uses should be categorised as either criterion-referenced or norm-referenced as he maintains "that proficiency decisions should be made on the basis of norm-referenced proficiency tests" (p.10). Brown's assertion of the clear cut distinction between the CRT/NRT is also evident in his presentation of the varying statistical formulas of validity and reliability suitable for each type. Similarly, Bachman (2004) reinforces the distinction between the two types by clarifying the statistical analysis procedures necessary for writing and analysing a norm and criterion-referenced test. Weir (1983) explains this distinction as follows:

> NR tests are designed and developed to maximize distinctions among individual test takers, which means that the items or parts of such tests will be selected according to how well they discriminate individuals who do well on the test as a whole from those who do poorly. CR tests on the other hand, are designed to be representative of specified levels of ability or domains of content, and the items or parts will be selected according to how adequately they represent these ability levels or content domains (p. 75).

Linn and Miller (2005) summarise the differences between CRT and NRT into four points as shown in the table below, arguing that the differences are only "a matter of emphasis".

Table 2.2. Differences between NRTs and CRTs according to (Linn & Miller, 2005, p.39)

NRTs	**CRTs**
Typically covers a *large* domain of learning tasks, with just a few items measuring each specific task.	Typically focuses on *delimited* domain of learning tasks, with a relatively large number of items measuring each specific task.
Emphasises *discrimination* among individuals in terms of their relative level of learning.	Emphasises *description* of what learning tasks individuals can and cannot perform.
Favours items of average difficulty and typically omits very easy and very hard items.	Matches item difficulty to the learning tasks, without altering item difficulty or omitting easy or hard items.
Interpretation requires a clearly defined group.	Interpretation requires a clearly defined and delimited achievement domain.

One of the early studies on using criterion-referenced language assessment is Hughes (1986). This study, in what the researcher described as an innovation in language testing, presented the case of developing and implementing a criterion-referenced English language test that was used as a gatekeeper to an English medium university in Turkey. The criterion-referenced test was "based directly on the English language skills that the students would need in their undergraduate studies" (p. 35). Hughes claims that one of the advantages of this type of test is the "washback" or "backwash" where the teachers teach towards mastering only the skills covered by the test. Nowadays, washback is viewed as a complicated consequence of assessment that can have either positive or negative outcomes; this will be discussed further in Section 3.3.4.1. Two other stressed advantages of criterion-referenced assessment are meeting "the information requirements of all stakeholders in the program" (Mckay and Brindley, 2007 p. 71), and relating more to classroom-based formative assessment (Rea-Dickins, 2007). Criterion-referenced assessment can be used to provide students with required feedback when using smaller units of assessment that focus on course outcomes. Also, by the end of a course, students would know what outcomes they have mastered and what outcomes they have not, from their scores. At the same time, scores in criterion-based assessment can be intrepreted in a conscise and clear summary of a list of points on attained outcomes for policy making purposes.

However, it has been claimed that in the classroom context the distinction between norm-referenced and criterion-referenced assessment tends to disappear. Banjeree and Wall (2007) found that in some cases teachers used a criterion-referenced checklist in a norm-referenced manner to compare the students' performances against each other. The fact that different assessment instruments may be categorised by criterion or norm-referenced (Martuza, 1977) depending on decisions that are usually made at administrative levels might contribute to the teachers' uncertainty about the appropriate contexts for using each type. Also, the fact that norm-referencing procedures are more 'well established' in educational contexts than the criterion referencing ones might explain why some teachers tend to use the former as the norm. Hughes (2003, pp. 21-22) says:

> books on language testing have tended to give advice which is more appropriate to norm referenced testing. One reason for this may be that the procedures for use with norm-referenced tests … are well established, while those for criterion-referenced tests are not.

Still, the reasons behind some teachers' confusion about norm and criterion-referenced procedures are not entirely clear.

2.2.3.3. Outcomes-Based Assessment: Validity and Politics

The discussion about the differences in formative/summative assessment and norm/criterion-referenced assessment cannot be complete without mentioning the increased use of outcomes-based assessment to serve summative and decision making purposes. Brindley (2001, p.393) defines outcomes based assessment as:

> systems that use pre-specified descriptions of learning outcomes – known, amongst other terms, as 'standards', 'benchmarks', 'competencies' and 'attainment targets' – as a basis for assessing and reporting learners' progress and achievement

Brindley (1998) argues that outcomes-based assessment is a response to the political pressures on educational systems to be more transparent and accountable; it reflects, he

emphasises, economic and market-oriented approaches. Political interference in language assessment seems to be widely recognised. Teasdale and Leung (2000) conclude that "it is also true that the wider political, social and ideological environment is likely to have an influence in decisions about the nature of assessment" (p. 180). It is argued that the rising interest in and use of outcomes-based assessment is far from being driven by educational motives only, or founded on educational principles solely. Brindley (1998) discusses the political and economic orientation of outcomes-based assessment:

> the widespread introduction of corporate mangement principles such as competion, productivity and cost-effectiveness into education has meant that educational policy and planning have become increasingly driven by considerations of economic accountability … while assessment and reporting mechanisims at the system level have become more outcomes-oriented, centralised and bureacratic to serve national economic goals, at the clasroom and local level the focus has shifted back to the individual learner (p.46).

This type of assessment is increasingly considered a tool for making high-stake decisions about students' language ability; however, many researchers have questioned its validity (e.g. Brindley, 2001; Llosa, 2007; Teasdale & Leung, 2000). To some educators outcomes-based assessment is purely formative and diagnostic; and it is not suitable for reporting students' overall learning achievement (e.g. Brindley, 2001; Torrance & Pryor, 1998). Brindley (2001, p. 398) discusses a number of concerns raised in the literature on the validity and reliability of teachers' assessment as a form of outcomes-based assessment, some of which were: low levels of generalisability, low reliability, inconsistencies in application interpretation of assessment criteria, and inconsistencies in transcription of language samples used as evidence of attainment. Besides, the issue of 'power' has been discussed in regard to outcomes-based assessment: it is argued that this type of assessment denies teachers autonomy in teaching and assessment. McKay and Brindley state that:

> first, teachers are given governance of process - that is the way that they will teach - but their decisions must be made within a context of externally mandated outcomes that must be achieved by all students. Second, because of this, teachers are inevitably divested of a certain amount of power in the assessment process (2007, p.73).

Moreover, several authors warn that many teachers and some adminstrators are underprepared for this type of assessment because it requires skills that are different from those for which teachers have normally been trained (Brindley, 2001; Teasdale & Leung, 2000). Likewise, Fox found that "many teachers and administrators are grossly under-prepared to carry out assessment agendas in either high- or low-stake contexts" (2008, p. 106).

This section has discussed different ways of categorising tests and assessment instruments based on their purposes, approaches and uses, the next part distinguishes between the qualities of standardised tests and performance assessment both of which were used in this study as a measurment tool for achievement in the FP English language courses.

Classifying tests as summative, formative, criterion-referenced or norm-referenced entails certain consquences with regard to how the results of these tests could be used or intrepreted. FP assessment, as will unfold in the following chapters, could be categorised as criterion-referenced that serves summative purposes. To understand the impact of these classifications, it is critical to identify the differences amongst these types.

2.3. Distinction between Standardised Tests and Performance Assessment

The history of testing in general goes back at least 2000 years to the Chinese Imperial examinations (Spolsky, 1990). Spolsky narrates that in the sixteenth century the Chinese examination system was brought to Europe and was used in the form of the Treviso test of mastery of curriculum in classical Christian schools. In 1853, in the British Parliament, Macaulay argued for using the Chinese examination system to select cadets for the Indian Civil Service (Spolsky, 2008). By the end of the nineteenth century, examinations found their ways into schools to measure students' achievement (ibid).

Though previous tests have included elements that measured candidates' language proficiency, Spolsky (2008) argues that the 1960s marked the beginning of standardised language tests as they are known today; other forms of language testing started earlier than that date. He says

> this, as they used to say in the old continuous movie houses, is sort of where we came in. For many current language testers, the history of our field seems to start in the 1960s, the beginning of large-scale industrialization and centralization of language testing that has come to be based in Princeton and Cambridge (p.447).

The methods used to assess language abilities were greatly affected by the changes in the theory of language conceptualisations and the inclusion of sociolinguistic aspects. In the last four decades, several models of the nature of language and its underlying competences and demonstrated performances have been developed (e.g., Hymes, 1972; Canale & Swain, 1980; Canale, 1983; Bachman, 1990). Fulcher and Davidson (2007) describe the relationship between these models and language tests, as follows:

> A model helps us to articulate the theoretical rationale for our test, and relate the meaning of specific test performance to language competence and ability for language use. Such models are constantly evolving and changing as our understanding of language acquisition and language use changes over time (p. 51).

The various models of what constitute language competence and performance have inevitably resulted in varying approaches and instruments used to evaluate what is believed to represent language abilities. There are at least three competing but different types of language assessment that dominate this field namely: psychometric testing, performance assessment and alternative assessment. Several writers agree that all these types are encompassed by the concept 'assessment' (e.g. Clapham, 2000; Lynch, 2001) however; they assert that they are distinct in multiple ways. Performance assessment in which specific tasks elicit some sort of language performance directly such as a speaking or writing tasks (Skehan, 2001) is linked to the 'direct testing' movement that spread in

the 1970s (Bachman,1990; 2002), and encompasses performance testing which elicits "performance that is available through the test setting" (McNamara, 1996, p.447). Performance assessment is sometimes considered as a type of alternative assessment and both are seen as distinct to traditional testing (Fox, 2008). Brown and Hudson (1998) differentiate between the three types of assessment, saying that performance assessment can be largely equated with 'constructed response assessment' whereas alternative assessment can be largely equated with 'personal-response assessment' and what they call 'tests' involve 'selected-response assessment', but can sometimes also include some constructed assessment tasks. Marking the distinction amongst performance assessment, tests and alternative assessment is significant in this study as the following discussion will reveal.

This study investigates 'tests' and 'performance assessment' which are both used in the Foundation Programme (FP): The term 'tests' will be used to refer to the General English Skills (GES) mid-term and final tests. These tests include tasks that require constructed performances (e.g. writing a short essay) in addition to indirect test tasks (e.g. multiple-choice or fill-in-the gap tasks). This use of the term *tests* follows Clapham's "construction and administration of formal or standardised tests" (2000, p.150). The terms *performance assessment* and *continuous assessment* will be used interchangeably to refer to assessment instruments in the Academic English Skills (AES) course which include writing a report and conducting a presentation. The term *continuous assessment* is the label used by CAS to refer to the AES assessment instruments, so it is more logical to use this to present data and discuss findings. The term *performance assessment* is used to facilitate linking the findings of this study to the literature in the field because the assessment instruments used in AES courses are encompassed by the term performance assessment in the field of language testing. This use of the term performance assessment conforms to Brown and Hudson's (1998) understanding of the term. They say that:

> performance assessments require students to accomplish approximations of real-life, authentic tasks, usually using the productive skills of speaking or

> writing but also using reading or writing or combining skills. Performance assessments can take many forms, including fairly traditional tasks like essay writing or interviews or more recent developments like problem-solving tasks, communicative pair-work tasks, role playing, and group discussions (p. 662).

Fulcher (2010, p.67) explains the different paradigms inherent in standardised testing and assessment saying "some believe that they [standardised assessment/tests and classroom assessment] are not only different in paradigms, but exist in a state of conflict". Lynch refers to assessment as being a broad term that includes both psychometric instruments as tests and non-psychometric instruments as writing a report.

> Assessment, in this conceptualization, is the superordinate term for a range of procedures that includes measurement and testing but it is not restricted to these forms. That is at times the systematic information we gather in order to make decisions about individuals comes from tests or other measurement procedures. At other times, however, we gather systematic information in a non-quantitative procedure, and we use that information to make decisions about individuals without quantifying it (Lynch, 2001, p. 358).

The distinction between testing and performance assessment is sometimes limited to how aspects of validity and reliability are considered in each type. Within this understanding, Clapham describes her view of tests and performance assessment saying: "I shall use the term 'testers' for those who concern themselves with the requirements of validity and reliability, and 'assessors' for those who are not consciously guided by such constraints" (2000, p. 150). The high profile of psychometric tests in the field of educational measurement and consequently language assessment is affected by some aspects of the popular psychological tests of intelligence in the United States, as Spolsky explains (1995). These tests usually employ psychometric measurements to maintain objectivity and increase reliability standards. Under the influence of psychometric measures, several authors in the field of language testing (e.g., Brown, 1995; Hughes, 1986) advocate conducting reliability and validity studies in both norm and criterion-referenced tests and urge test writers to consider both traits equally.

Clapham (2000) considers the stakes of the assessment as being another factor that distinguishes the two types; she explains that "there seems, indeed, to have been a shift in many language testers' perceptions so that they, perhaps subconsciously, may be starting to think of testing solely in relation to standardised, large-scale tests" (p. 150). In a similar vein, Davidson et al. (1997) claims that the differences between tests and performance assessment are often related to their high/low stakes.

> another source of distinction between 'tests' and 'assessments' is that some educators and applied linguists feel that high stakes' tests, which have a direct bearing on students' immediate future, need to have validity and reliability built into them, but that 'low stakes' tests such as classroom tests, which do not have such an obvious impact on students futures, do not (p.151).

Both Clapham and Davidson are clearly challenging these definitions and assumptions as rather simplistic. Fulcher (2010) has a slightly different view, but also issues a warning. He notes a tendency to use tests in high stakes context, but advocates using more formative type of assessment in the classroom.

> The technology of standardised testing has been developed in order to produce an engine that is capable of driving a meritocratic social system. Tests encourage learning because they are gateways to goals. In the classroom, however, we wish to devise engines that encourage learning, not only by motivating learners, but also by providing feedback on learning and achievement to both learners and teachers (p.67).

Some authors argue that the language skills sampled by tests or performance assessment constitute a third factor that discriminates between each method of assessment (e.g., McNamara, 2008), whilst others argue against such rigid borderlines; for example, Shohamy (1995, p.189) uses the terms "performance tests" and "performance assessment" interchangeably on a paper entitled *Performance Assessment*. She explains performance assessment as "tests where a test taker is tested on what s/he can do in the second language in situations similar to real life", clarifying that the term "tests" does

not entail 'real life' activities while the term "performance tests" does. Despite the occasional interchangeable use of the terms "test" and "assessment", generally they entail different epistemological considerations and distinct approaches to validity and reliability, the following two sections focus on clarifying the differences between these two methods of assessing language proficiency.

2.3.1. Epistemological Considerations

The increased use of performance assessment in language could be ascribed to the evolution in understanding the nature of language, as well as the epistemological positions on the nature of knowledge adopted by researchers in social sciences. The proponents of the first explanation argue that the realisation of language complexity had led language testers to the inclusion of assessment instruments other than traditional testing such as performance assessment (Spolsky, 2004). Expressing a similar understanding, Shohamy (1995, p. 195) states that the shift to performance assessment was driven by the movement towards creating real life (authentic) tasks; she adds that "communicative performance signifies the realization of the user's underlying communicative competence". She argues that the reasons for a dramatic increase in using performance assessment to evaluate language proficiency are eightfold (1) the blur in distinction between competence and performance; (2) the wide recognition of 'communicative performance' in the field of language teaching; (3) the narrow range of tasks used in existing tests; (4) external pressures to show face validity and that tests are "testing what they are expected to test"; (5) the effect of Hymes' views on communicative competence; (6) the trend towards communicative teaching strategies; (7) the widespread use of scales; and (8) the clear relationship between performance assessment and criteria used when performing needs analysis (Shohamy, 1995. pp.196-197). Messick in a similar vein claims that:

> performance assessments are becoming increasingly popular because they promise authentic and direct appraisals of educational competence leading to positive consequences for teaching and learning (1994, p. 13)

In the same vein, Clapham (2000) observes that this dialogue of paradigms has been paralleled by some alterations in some language testers' convictions about the nature of language and in their choices of assessment instruments. This has led some scholars to the view that traditional tests should be used in combination with or sometimes replaced by performance assessment. She argues that the emergence of performance assessment was "due to the influence of post modernism, that many 'testers' are rejecting the positivist principle that there is an independently existing reality that can be discovered (or measured) using objective, scientific method" (p. 151).

Other scholars attribute the change in language assessment focus from, predominantly using tests to increasingly using performance assessment in even high-stakes contexts, to the wider change in research epistemologies and views on the nature of knowledge. Just as positivism and constructivism have distinct epistemological underpinnings, so too do tests and performance assessment. Lynch (2001, p.362) explains this view:

> Testing as a measurement-driven enterprise is wedded to the current post-positivist research paradigm. It is centrally concerned with measuring, however imperfectly, traits and abilities. Underlying that research and practice are the assumptions that reality - in our case the reality of language and language use - exists independently of our attempts to understand it; that it is an objective entity that can be measured with proper tools and procedures. Alternative assessment, as an alternative paradigm, takes the view that language ability and use can best be understood as [a] realm of social life that does not exist independently of our attempts to know them. Judgments or decisions about language ability and use cannot, therefore, be accomplished as a measurement task: there is no 'true score' waiting to be approximated.

Likewise, Fox (2008, p.98) links "alternative assessment"/"performance assessment" and "tests" to opposing epistemological stances, claiming that these two genres in language assessment manifest different beliefs about the nature of knowledge and how it should be appropriately gauged, "testing culture is associated with positivist or post-positivist perspectives and assumptions" (Fox, 2008, p. 102) and explains further that tests, in the positivist view, represent language abilities as objective realities that should

be measured. On the other hand, alternative assessment and performance assessment, in the constructivist and socio-cultural view, see language abilities as constructed realties in a socio-cultural context. The shift to more inclusion of performance assessment represents a shift in beliefs. Gipps claims that performance assessment is a "change in view, and indeed the paradigm shift is part of the post-modern condition: a suspicion of belief in the absolute status of 'scientific' knowledge" (1994, p.288). Thus, the shift towards performance assessment relates to a shift towards understanding the nature of language as a communicative performance, and wider constructivist model behaviour.

2.3.2. Validity and Reliability Issues

The validity and reliability considerations for each type of assessment instruments are part of a fierce debate about their appropriateness, effectiveness and fairness of their use, particularly in high stakes contexts. Fox associates the epistemological stances of tests and performance assessment with how validity and reliability are considered. Referring to the proponents of performance assessment, she states that "some equate authenticity in alternative assessment with both reliability and validity … [this] perspective is deeply rooted within an interpretive or constructivist tradition, which views language as socially constructed and situated in contexts of use-rather than an underlying trait or ability which remains stable across contexts" (2008, p.101).

Proponents of performance/alternative assessment claim that the communicative nature of the tasks, that usually resemble real life tasks, enhances their validity to the extent that their low reliability might be overlooked (Gipps, 1994). Gipps argues that "if traditional test development has over-emphasised reliability at the expense of validity, performance assessment has in the same way over emphasised validity at the expense of reliability" (p.103). Gipps criticises the proponents of standardised tests for promoting reliability and norm-referencing, while "issues of validity and usefulness to teachers have sometimes been overridden or ignored" (1994. p7). She argues that the generalisability and reliability of performance assessment could be achieved by increasing the number of the tasks and making them versatile enough to represent the constructs tested. In the

same vein, Fox (2008) explains that performance assessment operates on a different understanding of what constitutes language, and consequently its definition of validity and reliability is understandably different. Unlike standardised tests that are founded on a psychometric basis and rigorously attend to validity and reliability issues, performance assessment is "rooted in an assessment culture rather than a testing or measurement culture" (Fox, 2008, p. 18). This culture promotes sharing power and celebrates fairness and equity by considering individual differences (Fox, 2008). Likewise, Rea-Dickins stresses that the "traditional and psychometric approaches are incompatible with the values underlying particular pedagogies and curricula" (2007, p267). Rea-Dickins, along with other authors (e.g., Gipps, 1994; Lynch, 2001), calls for a different understanding of validity and reliability that can be reconciled with the different principles of this type of assessment. Thus, these two qualities should be considered differently in designing, writing and analysing each type of assessment instrument.

On the other hand, some authors oppose using performance assessment in high stakes situations because of its low level of reliability; they criticise the attempt to redefine reliability to suit the epistemological position of this type of assessment (e.g., Teasdale& Leung, 2000). Likewise, some writers dispute the claim that performance assessment can be valid without implementing the psychometric measures that large-scale tests abide by (e.g., Brindley, 2001). They argue that language assessment should represent authentic language use without sacrificing the reliability of the instruments (Bachman & Palmer, 1996). Bachman (2002) maintains that "because of the complexity and diversity of tasks in most 'real life' domains, the evidence of content relevance and representativeness … is extremely difficult to provide" (p. 453).

Similarly, Teasdale and Leung criticise Gipps' view of validity in performance assessment as the characteristic of test or test writing; he says "validity appears to be narrowly conceived as a property of a procedure rather than of test scores … the claims of Gipps (1994), too, are difficult to reconcile with the kind of unitary approach to validity suggested in Messick (1989)" (2000, p.165). Performance assessment has also

been criticised for its low reliability as revealed by documented inconsistency and subjectivity in using rating scales (Clapham, 2000; Banjeree and Wall, 2006; Eckes, 2008). Similarly, Mckay and Brindley (2007, p.76) point out that "teachers [in their study] also tended to rely on observations based on their own 'intuitions', which do not necessarily mirror the assessment framework".

Furthermore, the practicality of using performance assessment has been described as problematic; "efficiency, comparability, and economy pose potentially formidable stumbling blocks for the implementation of a performance based examination system of the type being proposed" (Linn, 1993, p. 9). In this context, Linn recognises Gipps' argument of increasing the number of tasks as an essential and effective approach to deal with their lack of generalisability, but affirms that other challenges remain.

2.4. Combining Scores from Performance Assessment and Tests

Returning to the argument about the distinction between performance assessment and tests with regard to their validity considerations and epistemological stances, one might wonder if the scores from these two seemingly distinct types can be combined to represent a comprehensive picture of students' language proficiency and assist future decisions made about students' language proficiency. In CAS, students' scores in both the tests and continuous assessment (i.e. report writing and presenting) are combined; the total mark should exceed a cutoff point (i.e., 50%), if the students are to pass to Year 1. However, in the field of language testing, there are conflicting views about combining scores from multiple assessment instruments that seem distinct.

Teasdale and Leung (2000) call for separating the two types of assessment and considering their scores differently because of their different approaches to validity and reliability. He warns that the scores yielded by each type have different interpretations.

> In reality, the development and validation of the tests used seem to conform to standard psychometric practice; whereas the development and validation of teacher assessment have been neglected … the threat to validity when using such different assessment approaches does not appear to have been addressed. Consequently, both what is measured and, therefore, the meanings which can be ascribed to the scores are taken as unproblematic and somehow obvious (p.166).

However, some studies reject the claims that the students' results in tests and continuous assessment should not be combined. Llosa (2007), in a longitudinal study that lasted for three years, used a multivariate analytic approach to study the extent to which continuous assessment and standardised tests similarly or differently measured the same constructs. The study investigated a standard test used in a Californian school for fourth graders. The researcher found that the classroom assessment results were "consistent" with those of the standardised test and explained that both kinds of assessment had shared the same descriptors. She also claimed that the consistency of the results could be attributed to the time factor that increased the teachers' sense of what the descriptors used represented in terms of language levels. However, she found that the results in continuous assessment showed minimal discrimination between the skills, and attributed this to the wording of the descriptors. Some of the descriptors integrated performance assessment of multiple skills, e.g. "use expanded vocabulary and descriptive words and paraphrasing for oral and written responses to texts" (p.510). A second possible cause for the lack of discrimination between the language skills, as the writer suggests, was the halo effect associated with the teachers who assessed all the skills/traits. The results of this study are in line with Nitko's (1995) suggestion that the results of continuous assessment and tests could be combined to yield a comprehensive image of a student's performance. Nonetheless, performance assessment's lack of reliability continues to be seen by some as a barrier to considering scores generated by performance assessment equal to scores generated by tests. This emphasis on psychometric measures is highlighted by many scholars, including Davies, who links the validity and reliability of assessment to ethicality and morality; he asserts that "being ethical in language testing could be guaranteed by the traditional prospects of reliability and validity" (2008, p.441).

As FP assessment consists of tests and performance assessment, all of the above debated issues are vital in understanding how this combination of two arguably distinct assessment tools are implemented and how they are perceived by FP students and teachers. Identifying the literature on the distinction amongst assessment instruments is useful in explaining some of the findings on how these instruments are used or viewed and how the students' performance in each of them correlate with their performance in the other.

2.5. Chapter Summary and Conclusion

In general, standardised tests and other assessment instruments are classified according to their purposes, approaches, uses and other qualities into different types. Most of these types are applicable to all assessment instruments including tests. Assessment types not only differ in their purposes and uses, but also in how they are constructed and analysed. For example, the statistical procedures used in analysing scores in norm-referencing tests are different from those used to analyse scores in criterion-referencing tests. Also, test tasks in a proficiency assessment are different from those in an achievement assessment. In this study, these differences are highlighted when the assessment instruments are analysed and the implication of these differences are discussed.

Standardised tests and performance/alternative/continuous assessment differ intrinsically in two main ways: epistemological considerations and approaches to validity and reliability. These differences are not as clear though between performance, alternative and continuous assessment (Fox, 2008). In this Study, therefore, both terms 'performance assessment' and 'continuous assessment' are used interchangeably to refer to the assessment tasks used in the AES course which includes report writing and presenting.

Chapter 3: Language Assessment Validation and Programme Evaluation

"[b] Broadly speaking, validity is nothing less than an evaluative summary of both the evidence for and the actual as well as potential consequences of score interpretation and use" (Messick, 1995, p.5)

3.1. Introduction:

This chapter discusses relevant literature on three widely researched topics: language programme evaluation, language assessment validation, and predictive validity of language assessment. These topics inform and guide the general framework of this study and provide the context for discussing the results presented in Chapter 11. This chapter covers what might seem to be two completely different topics, namely current conceptualisations of assessment validity, and programme evaluation. Section 3.2 displays the earlier and current understanding of validity and validation focusing on some proposed frameworks for undertaking validation studies. Section 3.3 examines some current arguments and views on programme evaluation, and covers definitions, types, approaches, and epistemological paradigms. In section 3.3.4, the interconnectedness between the premises of programme evaluation and assessment validation is argued. One theme that underlines both areas of research is the consequences of scores' use and interpretations generated by the process of evaluating and assessing a programme or students.

Section 3.4 examines the findings reported by previous studies on the predictive validity of language assessment instruments implemented in the context of higher education. These studies are divided into research about the predictive validity of IELTS, TOEFL and in-house assessment. In this section, methodological and non-linguistic factors that affect the predictive validity of these tests are discussed.

Through out the chapter, I have focused almost entirely on studies which (a) were published in the last 30 years, (b) are in English and about English teaching/tests, and (c) from journals or other sources aimed at an international audience. Such limitation is common practice and seems justified for coherence, but I recognise that ideas which seem new in this community of practice may sometimes have existed long ago and have antecedents elsewhere, usually under different names.

3.2. Assessment Validation

This section deals with some theoretical arguments about assessment validation. It starts with an exploration of the changes in the meaning of assessment validity, then, discusses suggested frameworks for the latest conceptualisation of validity.

3.2.1. The Meaning of Assessment Validity

For a long time, it has been generally considered that a good test should be reliable and valid. A test's reliability is shown if similar scores are obtained when the same test is administered to two groups, equal in ability, or administered to one group at different times (Hughes, 2003). Harrison says "the *reliability* of a test is its consistency" (1983, p.10, italics in original). Test validity has been mainly viewed as five separate validities (i.e., face, content, predictive, concurrent, and construct) that represent distinct psychometric characteristics of a test. Sometimes these validities are grouped into internal, external and construct validities. The internal validity consists of face validity and content validity, whereas the external or criterion validity (Martuza, 1977) consists of concurrent validity and predictive validity. Hughes (1986, pp.22-28) explains the meaning of each type, saying that face validity signifies that an assessment looks suitable for its purposes; content validity means that an assessment is representative of the skills and content which it is supposed to measure; concurrent validity is established when an assessment correlates well with another test that similarly assesses the same constructs undertaken at about the same time; predictive validity means the extent to which an assessment predicts future performance of assessed participants; and construct

validity indicates that an assessment instrument measures the skills and abilities (i.e., constructs) that it is supposed to be measuring. This view sees reliability as a distinct quality from validity but both are necessary for a good test.

This view was challenged by Messick's seminal article (1989) in which he not only redefined validity as a unitary concept that involved multiple facets, but also argued that the consequences of a test should be included as an aspect of validity. He affirmed that the consequences of a test constituted an inherent facet of any evaluative judgement of the "*adequacy* and *appropriateness* of *interpretations* and *actions* based on test scores" (Messick, 1995, p.5, italics in original). Validity was defined as "a unified though faceted concept", and validation as a "scientific enquiry into score meaning" (Messick, 1989, p.6). Test validity in this unitary understanding still consists of the former validities, but they are seen as aspects, not independent entities, and they are encompassed by the overarching construct validity which links evidence from all other aspects, including the novel consequential aspect, to constitute one comprehensive concept. Bachman (2004) clarifies the premises of validity in Messick's view saying that (a) validity indicates the quality of the interpretation not scores, (b) validity is a matter of a degree and is not static, (c) validity is specific to a particular use, and (d) validity involves a comprehensive evaluative judgment. In this view, test validation is presented as the process of collecting information that supports the appropriateness and correctness of the interpretations of the test scores (Messick, 1989; Bachman & Palmer, 1996; McNamara, 1996). Thus when a test is used for a purpose that it was not designed to fulfill, it becomes invalid (Baker, 1989). In this conceptualisation, Messick identifies two threats to construct validity: *construct-under-representation*, which entails failure to include vital components of a certain construct, and *construct-irrelevant variance*, which arises from using irrelevant tasks to the construct and thus increasing the difficulty of the tests. The process of validating an assessment instrument should examine how far these threats have been dealt with and provide evidence for the claims made by its developers about the scores' interpretations. The main differences between the older

conceptualisation of validity and the currently most influential one, which was proposed by Messick, are captured by Chapelle (1999) in the table below.

Table 3.1. Summary of Contrasts between Former View and Messick's View of Validity, (Chapelle, 1999, p.258)

Older View	**Messick's View**
Validity was considered as a *characteristic of a test*: the extent to which a test measures what it is supposed to measure.	Validity is considered as *argument* concerning test interpretation and use: the extent to which test interpretations and use can be justified.
Reliability was seen as distinct from and a necessary *condition for validity.*	Reliability can be seen *as one type of validity evidence*.
Validity was often established through *correlations* of a test with other tests.	Validity is argued on the basis of a number of types of *rationales and evidence*, including the consequences of testing.
Construct validity was seen as one of *three types of validity* (the three validities were content, criterion-related, and construct).	Validity is a *unitary concept* with construct validity as central (content and criterion-related evidence can be used as evidence about construct validity).
Establishing validity was considered within the purview of *testing researchers* responsible for developing large-scale, high-stake tests.	Justifying the validity of test use is the responsibility of *all test users*.

The unified conceptualisation of validity is also central to Messick's (1994a, 1996) model for the validity of performance-based assessment. Messick criticised previous conceptualisations for viewing validity as separate entities, and excluding the social consequences of language assessment (Messick, 1994a). As an alternative, Messick proposes a unified understanding of validity for performance assessment that encompassed six facets namely: content, substantive, structural, external, generalisability, and consequential. Like the theory on language test validity, all of these facets must be considered as part of the unified construct validity of performance assessment. Each of these facets is concerned with a certain aspect of language assessment that requires an appropriate type of evidence. For example, the content validity aspect addresses the representation and relatedness of content; the substantive aspect focuses on the "theoretical rationales for the observed consistency in test responses"; the structural aspect evaluates representativeness of the assessment tasks' scores; the generalisability aspect looks into how generalisable the meaning of the

assessment scores could be to other domains or tasks; and the consequential aspect examines the implications of the interpretations and use of assessment including issues of bias or unfairness (Messick, 1996, p.6).

Messick's theory of validity and validating language tests and performance assessment has been widely influential in the language assessment field and has been recognised by leading authors (e.g., Bachman, 1990; Alderson, Clapham, & Wall, 1995; Davies & Elder, 2005; Kane, 1992; Weir, 2005). In general, most of these authors welcome the inclusion of assessment consequences as an integral part of validity. Norris (2008, p.48) states that "even those voices opposed to including consequences within definitions of validity agreed that the consequences of test use must be evaluated". Some authors welcome the fact that this conceptualisation of validity puts diverse methods at the validators' disposal, unlike the previous conceptualisation of validity that was dominated by a positivistic view of validation and heavily emphasised correlation studies (Chapelle, 1999). Other authors have fully adopted this unified concept of validity: for instance, Weir (2005, p.13) defined validity as:

> a multifaceted and different types of evidence are needed to support any claims for the validity of scores on a test. These are not alternatives but complementary aspects of an evidential basis for test interpretation

Weir also quotes statements in support of the unitary concept of validity such as Bachman (1990) who states that "it is important to recognise that none of these [validity aspects] by itself is sufficient to demonstrate the validity of a particular interpretation or use of test scores". Likewise, Alderson, Clapham and Wall (1995, p.171) quote Bachman's (1990) definition of validity and assert that "these 'types' are in reality different 'methods' of assessing validity". All this suggests that the unitary concept of validity proposed by Messick is now widespread in the field of language assessment.

This unitary concept of validity has not only been theoretically acknowledged, but has also been influential in empirical work. For instance, Powers, Schedl, Leung, & Butler,

(1999) investigated the predictive validity of a speaking test by correlating international students' scores with evaluations of the speaking skills of native-speaking undergraduate students. The authors claimed that the design of this study considered the consequence aspect of validity since it explored how the test scores might represent students' actual speaking levels by comparing these scores with undergraduate students' evaluations of the same performances. They argued that the "undergraduate students were selected as evaluators because they more than most other groups, are likely to interact with TSE examinees" (Powers et al., 1999, p. 399). This study considered the consequences of test score interpretations by comparing them to native-speakers' evaluations. It found a strong correlation between the speaking test scores and the undergraduate students' evaluations of the same performances.

3.2.2. Frameworks for Language Assessment Validation

Critics of Messick's proposal have pointed to the tensions between the consequential and evidential basis of test validity (e.g., Markus, 1998) and the alleged unwieldy and impractical nature of assessment validation when using the unitary model (e.g., Kane, 1992). One major problematic issue is how to present data from different validity aspects in a unitary argument. Davies (2012) maintains that "it remains unclear how to combine into a coherent whole the different insights offered by validation: in other words, how to give the multiple resources of information that validation provides the unity that validity needs" (Davies & Elder 2005, p. 8, cited in Davies, 2012). Davies (2012, p.41) also underlines the difficulty of operationalising Messick's theory of validity, and maintains that "validity is regarded by Messick as essentially an empty concept, leaving all the heavy lifting to validation". He stresses that this difficulty specifically faces those who attempt to create practical frameworks using Messick's conceptualisation, and observed that "Kane also seems to have found it difficult to discuss validation in concrete terms" (2012, p.41).

A number of frameworks have been proposed as practical applications of unified validation theory by leading scholars, three of these frameworks are now exmained in more detail. Kane (1992) suggested using the Argument-Based Approach (ABA) as a framework or a "technology" to identfy the procedures through which the validation of an assessment tool could be realised. He differentiated between two approaches: one is based on observable attributes and the other on theoretical constructs. The former entails building arguments based on observable average performances of possible tasks without an explicit reference to theories, while the latter starts from theories and is realised by using an index of possible observable attributes. About two decades later, Kane reasserted the plausibility and practicality of the ABA framework and explained its two steps as follows: "First, specify the proposed interpretations and uses of the scores in some detail. Second, evaluate the overall plausibility of the proposed interpretations and uses" (2011, p. 4). In this framework, the evidence collected for validation arguments varies according to the intended test interpretation and uses, and both the evidential and consequential basis of test validity are central to the validation argument. The social aspect of validation in Kane's framework is apparent in the emphasis given to evaluating the social consequences and being accountable for any negative ones; he thinks that:

> the test users have the primary responsibility for addressing the social consequences of assessment program; we are all responsible for our actions … test publishers are responsible for the claims they make about how their tests can be used and about the benefits claimed for such uses (Kane, 2011, p.14).

In this framework, three criteria for evaluating arguments are proposed, namely clarity of argument, coherence of argument and plausability of inferences and assumptions. All in all, Kane's framework is consistent with the principles of validity and validation proposed by Messick; it focuses on gathering evidence about the intrepretations of the scores, formulating a unified argument and considering the consquential aspect of the test scores.

Weir (2005), in conformity with Messick's conceptualisation of evidence-based validation of tests scores' interpretations and their consquential implications, has proposed a framework which is slightly different from Kane's (1992, 2010). It involves collecting test validity evidence in two phases: before the test "a priori validity evidence" and after the test "a posteriori validity evidence". The a priori evidence includes two types: (1) theory-based validity, which is concerned with establishing a connection between the test tasks and their theoretical underpinnings via statistical analysis, and (2) evidence related to context validity, which refers to the test tasks' representation of the larger pool of tasks that the test should sample. Weir described this step as a necessary one to avoid the 'construct under-representation' or 'construct irrelevance' discussed above. The a posteriori validity is concerned with gathering evidence for scoring validity, criterion-related validity and consequential validity. The scoring validity is "the degree to which examination marks are free from errors of measurement" (2005, p.23), or what used to be known as "test reliability". The criterion-related validity includes concurent validity, 'a criterion which we believe is also an indicator of the same ability being tested' (Bachman 1990, p.248, cited in Weir, 2005); and predictive validity, which concerns a tests' ability to inform about future performance (predictive validity is discussed in more details in Section 3.4). The a posteriori validation also includes gathering evidence related to Messick's notion of consquential validity, which explores issues such as differential validity, washback and effects on individuals within a society (this is discussed in more detail in Section 3.3.4). Weir's comments on the current emphasis on consequential validity in the language assessment field and explains it as being a result of the shift from formative assessment to summative assessment, which has been primarily driven by policy makers.

In this framework, the validation argument is based on investigating pre-specified aspects of validities instead of letting the claims or interpretations of the scores lead the validation process and determine the types of evidence needed. It could be argued that Weir's framework merely reconstructs the elements suggested in the earlier conceptualisation of test validation and reorganises them into two phases. In fact, a close

investigation of this framework reveals the old trilogy of validity (i.e., construct, content and criterion-related), which are renamed as theory-based, context, and criterion-related validities respectively. Likewise, reliability is relabeled as scoring validity. It, nonetheless, incorporates an element of consquential validity. Though this framework emphasises the unity of these validities and the crucial role played by the consquential aspect, it ignores the notion of test uses and interpretation which are central to Messick's theory of validity. For example, Weir (2005, p.44) presents a framework for validating reading, listening writing and speaking tests that links context, theory based, scoring, consquential and criterion-related validities to test-takers' charicteristics but not to the test uses or score interpretations in the process of validation.

The third framework for evaluating and validating language assessment is the Assessment Use Argument (AUA) devised by Bachman and Palmer (2010). It is similar in principal to Kane's framework, but offers more detailed procedures for constructing a validation argument. AUA entails:

> a theoretically grounded and systematic set of principles and procedures for developing and using language tests and an understanding that will enable readers to make their own judgements and decisions about either selecting, modifying, or developing a language assessment whose use can be justified to stakeholders (Bachman & Palmer, 2010, p.30).

It is also described as a tool which test developers can use to show the validity of their tests. It "thus provides a framework for investigating the extent to which the intended use of a particular assessment is, in fact, justified" (p.32). Like Kane's framework, the AUA includes two procedures: articulating test uses and gathering relevant supporting evidence. However, the AUA is more detailed than Kane's framework in that it identifies six sources of information and how they are linked to each other using Toulmin's (2003) structure of practical reasoning. These are: data, claims, warrants, backing, rebuttals, and rebuttals backing. Bachman and Palmer (2010) explain that the *claims* are conclusions based on certain *data* or facts, and *warrants* were "propositions to justify claims" (p.96) that could be supported by *backings*. The *rebuttals* are

"conditions under which the warrant may not apply" (p.97) that could be reinforced by the backings. Test validation, according to this framework, starts from the claims made about the interpretations of test scores and works its way back to the qualities of the test by collecting evidence that supports stated claims using warrants, backings and rebuttals. Like the other two, in this framework both the evidential and consquential basis of test validty are considered.

Clearly, these three different conceptualisations of the unified notion of validity generate different arguments. Kane's and Bachman and Palmer's frameworks for assessment validation both focus on justifying claims about test uses and score interpretations; they could be described as selective and flexible as they build the validation argument on stated claims and gather only necessary evidence to support or contradict these claims. However, Weir's framework focuses on gathering a priori and a posteriori evidence needed for validity arguments, and is much more comprehensive and rigid as it approaches validity arguments using fixed procedures that look into specific aspects of validity regardless of the test written claims about score intrepretations or test uses.

Using these frameworks alone is not enough to answer the study questions which investigate multiple aspects of FP assessment: predictive validity, effectiveness, and impact. Using approaches from programme evaluation and validation studies is considered more suitable for the purposes of this study as will be explained below.

3.3. Programme Evaluation and Language Assessment Validation

Some authors in the field of language assessment validation have proposed recourse to programme evaluation methods and approaches to overcome challenges of current validation frameworks (e.g., Norris, 2008). Before elaborating on current approaches to programme evaluation, it is crucial to address an ambiguity that is common in the fields of programme evaluation and assessment validation. There seems to be a persistent haziness between the three terms *tests, assessment,* and *evaluation* which writers on programme evaluation in general and on language assessment in specific regularly

attempt to clarify (e.g., Bachman, 1990; 2004; Bachman and Palmer 2010; Lynch, 1996; Scriven, 2003). Bachman (1990, pp. 6-7), for example, differentiates between the terms *assessment* and *evaluation* by identifying the former as "the process of collecting information about something that we are interested in, according to procedures that are systematic and substantially grounded", and the latter as "one possible use of assessment" which "involves making value judgments and decisions on the basis of the information" (p.9). He argues that the terms assessment, measurement and tests should be used to signify tools that are used for evaluation purposes. In the same vein, Norris (2006) warns of confusing the three terms evaluation, measurement and assessment. Like Bachman, he argues that assessment is a tool for evaluation and states that "couched within this evaluative framework, assessment offers an important contribution to our understanding and improvement of what we do as college educators" (p.581). Thus, evaluation encompasses and utilises different assessment tools (e.g., reports, presentations, or portfolios) and tests measure a certain construct and generate an evaluative judgement.

3.3.1. Programme Evaluation

The assumption, however, that the process of evaluation must generate evaluative judgment in all contexts is sometimes argued to be false. For example, Scriven (2003) claims that the "tendency to refer to evaluation as essentially involving value judgments is mistaken and misleading" because it oversimplifies the concept, which might or might not include making "judgments of value". He adds that evaluative claims could be a result of simple and direct observation; for example, an evaluation of certain performances on easy tasks is direct and obvious and does not entail "weighing" or "balancing" (Scriven, 2003, p.28). Scriven generalises the meaning of evaluation to denote a process leading "to a *particular type* of conclusion - one about merit, worth, or significance - usually expressed in the language of good/bad, better/worse, well/ill, elegantly/poorly" (emphasis in original, p.16). The language of evaluation, he maintains, usually relies heavily on its context to reach evaluative conclusions. The prominence of

the context in evaluation studies is similarly underlined by Norris (2009), who affirms that "in their reports, the evaluators provide practical discussions of the background, contexts, stakeholders, and methods for evaluations, thereby situating and clarifying the distinct forms that contemporary language programme evaluation may take" (p.8). The following sections will first discuss the types and purposes of programme evaluation, then describe the evolution of its epistemological underpinning, and finally discuss how programme evaluation approaches can be utilised in validation studies.

3.3.2. Types and Purposes of Programme Evaluation

There are different purposes for conducting a language programme evaluation, the most common of which is responding to administrative requirements for accountability and decision making about the continuation of a certain programme. There can, however, be other purposes such as "generating theories, making policies, and improving professional practices" (Kiely, 2009). Rea-Dickins (1994) adds two other purposes in the context of language programme evaluation: enhancing curriculum and contributing to teacher development.

To meet these purposes and several others, a number of evaluation types, which vary in their questions, foci and methods have been devised. Writers in the fields of educational programme evaluation, (e.g., Owen, 2007; Scriven, 2003) and language programme evaluation (e.g., Kiely, 2001; 2009; Lynch, 1996) identify several types of evaluations. The ones suggested for use in language programme evaluation are largely derived from the ones used in the field of educational programme evaluation (Kiely, 2001), and share similar purposes and methods.

Owen (2007, pp. 41-51) identifies five "forms" of educational programme evaluation, some of which are replicated in the literature of language programme evaluation. The first form is *proactive* evaluation which aims at synthesising information about programmes and assumes that "what is already known should influence action". The

second form is *clarificative* evaluation which, as its name suggests, aims at providing clarification and involves the notion of making a programme's rationale and design explicit. The third form is *interactive* evaluation, which targets improvement and requires that information should be provided for stakeholders to achieve intended improvements. The fourth form is *monitoring* evaluation, which involves checking and refining programmes to ensure their quality by closely monitoring them. The fifth is *impact* evaluation, which focuses on measuring attainment of outcomes and accountability of a programme through identifying what 'works' and why.

In the field of language programme evaluation, Kiely (2001, p.243) suggests three comparable types of evaluation which have similar purposes and focal points to the ones discussed by Owen (2007). These are *functional* evaluation, which involves "delineating, obtaining, and providing useful information for judging decision alternatives", *developmental* evaluation that "celebrates the potential of evaluation for shared development of the program" and *critical* evaluation which involves "collaborative development, based on the power structures within programes" (Kiely, 2001). It is clear from these partial definitions that the *functional, developmental and critical* evaluations suggested for use in language programme evaluation share several features with the *claricative, interactive* and *impact* evaluations used in the wider field of educational programme evaluation. Evaluation purposes are not the only factors affecting how different forms of evaluations are produced; the epistimological premises considered in a programme evaluation are another factor in shaping the parameters of an evaluation study, and this is further explored in the following section.

3.3.3. Epistemological Paradigms in Programme Evaluation

In an early modern period of programme evaluation in educational contexts, positivist assumptions and approaches were dominant. Beretta (1992), in a review of language programme evaluations that were published between 1960 and 1985, reports that positivist quasi-experimental designs were the dominant type. Tests were often utilised

as the sole indicator of the effectiveness and appropriateness of a programme (Kiely, 2001; Lynch, 1996; Norris, 2008; Rea-Dickins, 1994; Scriven, 2003). Currently, however, a wide variety of epistemologies such as objectivism[6] and constructivism[7], interpretivism[8] … etc. can be traced in the design, implementation, and analysis stages of evaluation studies in the educational field. Lynch (1996) claims that the approaches used in language programme evaluations belonged to two epistemological views: positivist and naturalist, he (p.13) uses the term naturalistic "to include the alternative paradigms [to positivism and postpositivism], including constructivism and critical theory", though he acknowledges that naturalistic was re-labelled as constructivism by Guba and Lincoln (1989). Agreeing with other writers (e.g., Rea-Dickins, 1997), Lynch describes positivism as the dominant view in programme evaluation for decades and states that "in applied linguistics, in general, and in programme evaluation in particular, there has been a strong tendency to favor a traditonal, quantatitive experimental approach to conducting researh" (1996, p.13).

Programme evaluations following a naturalistic view considers what is evaluated as a continuously changing process. When following a positivist view, it considers what is evaluated as a stable, fixed and unchanging fact (ibid). Thus it is vital in conducting programme evaluations that the epistemological paradigms are stated.

3.3.4. Bringing Together Assessment Validation and Programme Evaluation

In the last three decades, approaches from the field of educational programme evaluation have been brought to the forefront of language programme evaluation research (Lynch, 1996). Some claim that evaluation of language programmes started in the 1950s when

[6] Objectivism is an epistemology under which various theoretical perspectives situate such as positivism. In discussing positivism, Crotty (2009, p.29) states "this supreme confidence in science stems from a conviction that scientific knowledge is both accurate and certain. In this respect scientific knowledge contrasts sharply with options, beliefs, feelings and assumptions that we gain in none-scientific ways".
[7] "What constructionism claims is that meanings are constructed by human beings as they are engaging with the world they are interpreting" (Crotty, 2009, p. 43).
[8] "interpretivism emerged in contradistinction to positivism in attempts to understand and explain human and social reality" (Crotty, 2009, p.66).

the results of the tests were used as an indication of the effectiveness of a programme (e.g., Rea-Dickins, 1994; Lynch, 1996). This view of programme evaluation in language assessment has been criticised for adopting a limited view of evaluation in considering test scores as the only measure for the effectiveness of a programme. Not only the paradigms, methods and forms of educational programme evaluation have been adopted in language assessment evaluations, but also a similar path of epistemological evolution has been followed (see Section 2.3.1).

Using programme evaluation as the framework for assessment validation is analogous to working on a mosaic: the small pieces are meaningless without a portrait or sketch of the whole mosaic that is built out of the pieces. Lynch (1996) suggests that investigating language programme assessment validation through the lenses of established evaluation approaches in the field of education adds an educational prospective to what has long been considered as a psychometric field. Norris notes that

> validation endures as the principal mandated and perceived requirement for the implementation and perpetuation of assessments within education; however, it is precisely *within* educational settings that the value of validation, the functional scope of its ends, and the suitability of its means, remain indeterminate (2008, p.72).

He also criticises the failure to use established frameworks and approaches of programme evaluation in assessment validation studies, in spite of the evident relatedness between the two areas of inquiry, Norris (2008, p.73) adds:

> it is apparent that current approaches to assessment validation resonate little with contemporary consensus on programme evaluation, this despite persistent (if not necessarily consistent) recourse within educational measurement rhetoric to the idea of evaluation as fundamental to the nature of validation and what it seeks to achieve.

One of the advantages of using programme evaluation approaches is that they could provide assessment validation with concrete and structured approaches that the field of

assessment validation currently lacks. Though the theoretical models of language *programme* evaluation have adapted models originating in the general field of programme evaluation (Lynch 1996), language *assessment* validation is still struggling to find practical frameworks for applying the unified validity theory, as discussed earlier (see Section 3.2.2). Both *programme* evaluation and *assessment* validation aim at reaching some sort of conclusion about the functionality of using certain tools, and share the concept of building coherent and comprehensive arguments about programmes' functionality. A second advantage is that programme evaluation can bring to the validation process the involvement of the stakeholders who might participate in assessment programme development (Norris, 2008). A third advantage is that programme evaluation considers assessment as a programme rather than discrete or independent tools (ibid). This view broadens the scope of assessment validation and allows for a comprehensive investigation of the effectiveness of assessment and its consequences.

As mentioned earlier, there are several types of programme evaluations that provide frameworks and approaches suitable for particular purposes of evaluators and natures of evaluand (e.g., Lynch, 1996; Owen, 2007; Scriven, 2003). One of the types of evaluations that correspond with the focus of this study is *impact* evaluation, which will be briefly introduced in the following paragraphs. Owen (2007) characterises impact evaluation as having "a strong summative emphasis in that it provides findings from which a judgment of the worth of a programme can be made" (p.252). He pinpoints four major areas with which this type of evaluation is concerned: (1) identifying a programme's outcomes, (2) examining correspondence between plans and implementation and how outcomes are influenced by implementations, (3) presenting evidence to stakeholders on how programme resources have been used, (4) providing information for future changes. Owen (pp.255-263) also lists six possible approaches that can be utilised in this type of evaluation, as follows:

- objective-based (i.e., investigates if the stated objectives of a programme have been met). The success of a programme is measured by achieving its set

objectives. Though this method was opposed at its early stages, currently it is widely used in both private and government sectors for management and appraisal purposes.

- needs based (i.e., examines if a programme is fulfilling a certain need). In reaction to an objective-based approach, this approach was first proposed by Scriven (1972) to include the needs of a programme in an evaluation as not all needs are necessarily included in a programme's objectives.
- goal-free (i.e., determines intended and unintended programme outcomes despite predefined programme objectives). In a stronger reaction to an objective-based approach, the goal-free approach was created. This rarely used approach proposes to examine all programme affects and intentionally overlooks specified objectives.
- process-outcome studies (i.e., investigating the extent to which a programme has been implemented). It concerns examining the implementation of objectives to understand the outcomes and uses methods such as: observations, interviews, self-reports and records to determine how much of the planned programme is implemented.
- realistic evaluation (i.e., examines a programme within its context to which the generalisability of its outcomes is limited). This approach is based on two premises: the first is that the results should be reached through an inquiry; the second is that in a social context the findings cannot be generalisable and are true for a specific context only.
- performance audit (i.e., entails objectively examining the degree of matching between claims and established criteria). The concept of 'performance auditing' has been adapted from accounting and it involves an evaluation of all activities related to a certain area to produce some recommendations.

These approaches employ various methods to tackle the different types of data. Data sources can include policy documents, programme statements, interviews, or score records as required by the purposes of the evaluation. The focus of impact evaluation

corresponds with the focus of validation studies and the approaches and methods it uses provide a structure that can be implemented in validation studies.

In this study, some approaches from impact evaluation (i.e., objective-based, needs based, goal free, process-outcome approaches) were utilized in investigating the effectiveness of FP assessment. Approaches from validation frameworks were used in investigating the predictive validity and content validity of FP assessment.

3.3.5. Test Impact and Consequential Validity

Both programme evaluation and assessment validation place considerable emphasis on the significance of the social context and consequences of any validation or evaluation (Bhola, 2003; Messick, 1989; Owen, 2007). In assessment validation, Messick makes an argument for including the consequences of an assessment in its validation.

> The questions are whether the potential and actual social consequences of tests interpretation and use are not only supportive of the intended testing purposes, but at the same time are consistent with other social values. Because the values served in intended and unintended outcomes of test interpretations and test use both derive from and contribute to the meaning of test scores, the appraisal of social consequences of testing is also seen to be subsumed as an aspect of construct validity. (1989, p. 18)

Some scholars argue that test consequences have not been appropriately addressed in practice and call for change in this area. Hamp-Lyons (1997), for example, disapproves of the situation in which the current international English language proficiency tests such as IELTS and TOEFL are not open to public feedback. She claims that this conflicts with postmodernist values and current thinking on language testing ethics. Likewise, Bachman and Palmer argue for attending to both positive and negative consequences: “in addition to the beneficial consequences we intend to bring about, we will also need to consider the potential unintended detrimental consequences” (2010, p. 87). In the light of this increased attention (though often only theoretical) to consequences in both validation and evaluation studies, the following sections discuss

three general topics within assessment consequences: washback, stakeholders and policy making.

3.3.5.1. Washback

Washback or back-wash, in its simplest definitions, refers to the impact of a test on learning and teaching (Wall & Alderson, 1993; Shohamy, 1996). Though some authors associate washback with negative effects on teaching and learning, empirical studies have found that washback is a complex consquence of tests that can be both positive or negative depending on several factors, such as the nature of textbooks, the language tested, future implications of test scores and research methodolgy (Hamp-Lyons, 1997; Shohamy, 1996; Wall & Alderson, 1993; Weir, 2005). Negative washback is associated with both language courses (e.g., Shohamy, 2001b) and non-language courses where teachers intentionally teach only the skills required by a test to improve students' scores (Hamp-Lyons, 1997). In this study, students and teachers were asked about their views on the influence of FP assessment on FP teaching and learning processes.

The manner in which washback affects students' performances in tests is not one-directional or straightforward. Green (2007) found that certain presessional courses that concentrated on teaching students the writing skills tested by IELTS did not result in significantly better scores compared to other courses. He compared participants' entry scores in the IELTS writing test with their exit scores in the same skill in three different courses. The first course was an IELTS preparation course, the second one general English for Academic Purposes (EAP) course, and the third a combination of the previous two. He reported that "there was no significant difference in terms of score gains between those studying on pre-sessional EAP programmes and those engaging in dedicated IELTS preparation courses" (p. 93).

3.3.5.2. Political and Policy Making Consequences

Messick (1996) emphasises the complexity of the political aspect of language assessment validation; he argues that "as a salient social value, validity assumes both a scientific and a political role that can by no means be fulfilled by a simple correlation coefficient between test scores and purported criterion … or by expert judgments that test content is relevant to the proposed test use" (p.5). This argument calls for a different conceptualisation of validity and different approaches to evaluating its political and social impact; it calls for approaches that go beyond the usual methods of validation using statistical correlation or expert evaluations; it seems to encourage using multiple methods that consider and attempt to gauge the social and political ramifications of language assessment or what is called the consequential basis of test validity.

Shohamy (2007) widens the term 'test consequences' suggested by Messick (1989) to include the impact which language assessment instruments might have on national language policies. She argues that tests are sometimes used to overtly or covertly enforce specific political agendas and "serve as de facto policies that can override and contradict existing policies and create alternative policy realities "(Shohamy, 2007, p.120). Other authors take a similar position, and claim that tests are not only used to influence language policies, but also political decisions about educational policies (e.g., Grek, 2009; Nunan, 2003; Spolsky, 2004). In this study, educational policies on language assessment constituted a major part of document analysis in which approaches from impact evaluation were used.

3.3.5.3. Stakeholders

Hamp-Lyons (2000) stresses that language assessment should consider the purposes of all the different stakeholders, adding that "under the influence of postmodernism, we cannot avoid acknowledging the contingent nature of knowledge nor the fact that different stakeholder groups will have competing goals and values" (2000, p. 581). She underlines the role of language testing in social and educational reforms and emphasises the consequent responsibility of language testers to the stakeholder. The need to accept

this responsibility has led to the creation of several codes of language testing ethics (e.g., Code of Ethics of the International Language Testing Association; ILTA, 2010). This responsibility towards the stakeholders is specifically evident in its first principle of this code.

> Language testers shall have respect for the humanity and dignity of each of their test takers. They shall provide them with the best possible professional consideration and shall respect all persons' needs, values and cultures in the provision of their language testing service (p.2).

The term stakeholders has been used to refer to students, parents, teachers, official bodies, and the marketplace (Hamp-Lyons, 1997; Rea-Dickins, 1997; Weir, 2005); and sometimes is expanded to include test writers, curriculum designers or policy makers (Alderson, Clapham, & Wall, 1995; Bachman & Palmer, 2010). Several empirical studies have investigated test impact on different stakeholders (e.g., Green, 2003); the majority of the studies, though, focus on the positive/negative impact a test has on teaching and learning or what is known as washback (e.g., Cheng, 1999; Hamp-Lyons, 1997; Messick, 1996; Shohamy 2001b).

Stakeholders' perceptions and understandings of the meaning of performance in language assessment constitute another facet of the consequential aspect of language assessment validity. The status and role of tests are influenced positively or negatively by how students and stakeholders perceive them (Fulcher, 1996). It has been argued that students' perception of knowledge is manipulated by what tests (or test writers) cover because the "right" knowledge is limited to what is considered a correct answer in a test. Fulcher (2010, p. 10) highlights this point by explaining Foucault's view on the power of examinations:

> For Foucault, the ritual is not a rite of passage, but a means of subjecting the test takers to the power of those who control the educational system. It is an act of observation, of surveillance, in which the test taker is subjected to the 'normalizing judgment' of those who expect compliance with the knowledge that is valued by the elite. After all, the answers that the test taker provides

> will be judged, and in order to do well they have to internalise what is considered ‘right’ by those in power.

Fulcher argues that tests are used as surveillance tools to normalise and control knowledge or what is “right”, and claims that the perception of those under control is systematised by tests.

Brown and Hirschfeld (2008) conducted a study in New Zealand schools about how students’ perception of assessment can affect their achievement. They provided 3,469 secondary school students with four different conceptions about assessment in a self-report inventory that were based on previous literature. They found out that “the two biggest and positive predictors of student achievement were school year … and the conception that assessment made students accountable" (Brown & Hirschfeld, 2008, p. 11). They also pointed out that "the students who enjoyed assessment experiences tended to assume that schools rather than students were being made accountable" (Brown & Hirschfeld, 2008, p. 11). They argued that the results of this study were in line with those conducted on self-regulation and formative assessment, in that students who perceived assessment as “taking responsibility” achieved better academic results than those who perceived assessment as an irrelevant, unfair, tool for improving learning, or enjoyable. Moreover, the authors indicated that the students’ perceptions of assessment were often shared by the teachers who had once been students themselves. One of the implications of this study for teacher training was a need to change teachers’ preconceptions of assessment in a way that emphasised their own accountability prior to sending them into the classroom.

Teachers' perceptions of assessment not only influence their performance and accountability, but also have an active role in shaping test design and marking. Knoch (2009) conducted a study on how 100 trained raters used two rating scales; one with less specific descriptors and the other with detailed descriptors. He found that that teachers were influenced by the "halo effect" especially when they faced difficulties in rating. In these cases teachers interpreted the scales globally or generally, this resulted in a

tendency to award the same score across more than one skill (Knoch, 2009). He reported that "the rating scale, and the way raters interpret the rating scale, represents the de-facto test construct" (p.276). This study and others on rater variability were discussed in Section 2.2.2.4.

3.4. The Predictive Validity of Assessment

Given that one of the purposes of the Oman General Foundation Programme (GFP) is to "prepare students for their postsecondary and higher education studies" (OAAA, 2009, p.4), and given that validation is a "scientific inquiry into score meaning" (Messick, 1989, p.6), this section reviews the findings reported by some studies on predictive validity of language assessment. We need predictive validity studies as one type of evidence towards verifying the claims and inferences made using test scores (Bachman, 1990; 2004; Bachman & Palmer 2010; Kane, 2010; Messick, 1995; Weir, 2005).

In spite of the widespread theoretical acceptance of the unitary view of validity that involves several 'aspects', studies on the predictive validity of language assessment are still carried out for their own merits (i.e. estimating future performance by correlating results on two different instruments separated by a specific time difference). Twenty-five years ago, Graham (1987) described the results obtained from predictive validity studies on language tests as inconsistent, and the same conclusion can be drawn today from the following selective summary which is divided into studies about the predictive validity of IELTS, TOEFL and in-house language tests.

3.4.1. Studies on the Predictive Validity of IELTS

Studies on predictive validity investigate "how well somebody will perform in the future" (Alderson, Clapham & Wall, 1995, p.180). Many studies have explored the predictive validity of IELTS in different contexts; the following table summarises the findings of some of those which are subsequently discussed.

Table 3.2. Some Studies on Predictive Validity of IELTS

Study	Country	Number of Participants	Type of Correlation	Correlation
Elder (1993)	Australia	32 international students	IELTS and administrators' ratings	0.5*
Cotton & Conrow (1998)	Australia	33 undergraduate & postgraduate students	IELTS and GPA	-0.24*
			ILTS and staff ratings	0. 15*
			IELTS and students' self-assessment	-0.28*
Huong (2001)	Australia	320 Vietnamese post and undergraduate students	IELTS and GPA	0.30*
Kerstjen & Nery (2000)	Australia	113	IELTS and GPA	Non-significant
Feast (2002)	Australia	101 international students	IELTS and GPA	0.39*
Woodrow, (2006)	Australia	62 students 15 teachers in Faculty of Education	IELTS and teachers evaluations	0.40*
Breeze & Miller, 2008	Spain	289 Spanish under-graduate students	IELTS and GPA in Humanities	0.34*
			in Law	0.28 **
			in Medicine	0.25*
Yen & Kuzma, (2009)	Britain	61 Chinese Business School students	IELTS and GPA	0.46**

* $p<0.05$
** $p<0.01$

Cotton and Conrow (1998) explored the predictive validity of IELTS in a study that included a sample of 33 undergraduate and post-graduate international students in an Australian University. To collect data, they utilised a seven point likert scale questionnaire and conducted interviews over two semesters with the students and some of the teaching staff. The students were from different countries, mainly Asian; and studied in different faculties (i.e., Engineering, Law, Health Sciences, Humanities, and Technology). The students' bands in IELTS were correlated with (1) their GPA[9] scores

[9] GPA stands for Grade Point Average which is a system for standardised measurement of achievement in a certain course. GPA is calculated by aggregating the grades a student earned in all courses divided by the total number of credit hours assigned to each course.

in both the first and second semesters, (2) academic staff ratings of students' performance, (3) and students' own assessment of their academic performance. The Pearson correlation coefficients were found to be r=-0.24 between IELTS bands and GPA, r=0. 15 between IELTS bands and staff ratings, r= -0.28 between IELTS bands and the students' ratings in the first semester, and r= 0.12 between IELTS bands and the students' ratings in the second semester. The researchers attributed the weak correlation to the high number of students in disciplines that were believed to depend less on students' language proficiency (e.g., Engineering). Another factor which the authors did not mention but which could have contributed to the low correlation was the inclusion of data from both postgraduate and undergraduate students from different disciplines. It has been suggested that the predictive validity of language assessment instruments is usually low in heterogeneous samples (e.g., Graham, 1987).

Elder's (1993) study, unlike most studies presented in this section, reported a strong correlation between students' scores in IELTS and their academic achievement. The study included 32 overseas students who were enrolled in a Diploma in Education programme; the course administrators' ratings of students' performances were collected as well as the students' IELTS scores and the students' responses on questionnaires about their perceptions of language difficulties. The correlation between students' scores in IELTS and administrators' ratings of the students was r = 0.5 in the first semester r = 0.14 in the second semester. The difference between the two semesters was explained by (1) the students' improvements in English language abilities, and (2) non-language variables whose effect was more likely to occur in the second semester more than the first. Elder (1993) concluded that the results were not conclusive but that they supported previous findings that IELTS was a better predictor at lower levels of language proficiency; and that above a certain level of language proficiency other factors played a more significant role in academic achievement. This is similar to Grahams' argument (1987) that there is a minimum level of language proficiency below which academic achievement may be more strongly affected by language weaknesses. This, however, is insufficient to explain the strong correlation that was found in the first semester between

the IELTS scores and administrators' ratings given that only very few studies have reported a similar high correlation. One explanation could be that, since the diploma in education included teaching in schools, the candidates' language proficiency played a major role in their conduct of lessons and their confidence as teachers and consequently the administrators' ratings.

Feast (2002) argues that the role of English language proficiency is important but not 'critical' to academic achievement. She maintains that English language is only one of several factors affecting academic achievement, and that there are other factors that could impact students' academic achievement such as: 'personal background' (i.e., age, gender, personality), 'academic background' (i.e., previous studies), culture, teaching or other types of support. Her sample included 101 international students from different specialisations at an Australian University. Using a multilevel regression, the association between students' scores in IELTS and GPA (an accumulation of course grades in up to five semesters) was measured. The results showed a "significant but weak relationship between English language proficiency, as measured by the IELTS scores, of international students and their performance, as measured by their GPA" (Feast, 2002, p. 83).

Another study that investigated the predictive validity of IELTS is Huong (2001) which focused on Vietnamese students in Australian universities who were admitted in the period from 1993 to 1999. The sample included 320 post-graduate and undergraduate students. The students' scores in IELTS and their GPAs in the first and second semester of university study were correlated using Pearson Product-Moment. The correlation coefficients for the IELTS and semester 1 and 2 GPAs were found to be $r= 0.33$, $p<0.05$ and $r=0.30$, $p<0.05$ respectively, in a similar range to most previous studies. Huong argued that the more homogeneous the sample was the higher the correlation between the IELTS scores and GPA tended to be. He discussed three factors that could have contributed to the inconclusiveness of the results about predictive validity among the previous studies: (1) using different measures of academic success, (2) the heterogeneity

of the samples, (3) and combining all disciplines together whether they were linguistically demanding or not. Though the results of this study were similar to many conducted on the same topic, the methodology of this study is questionable because of the small number of participants that belonged to each university/specialisation, and the varying grading systems used in the universities to calculate a semester GPA.

Two other studies that focused on the predictive validity of IELTS found similar results to Huong's. The first (Woodrow, 2006) utilised student questionnaires (*n*=62), teacher evaluations (*n*=15), and teacher/student interviews conducted in the Faculty of Education at an Australian University. The reported results revealed a moderate significant correlation ($r= 0.40$, $p < 0.01$, $n= 62$). The second study (Yen & Kuzma, 2009), investigated the predictive validity of IELTS scores for 61 Chinese students who were admitted to Worcester Business School, UK. The students' GPA in semesters 1 and 2 were correlated with their scores in EILTS; the correlation coefficient was $r=0.46$, $p<0.01$ for the first semester and $r= 0.25$, $p<0.05$ for the second semester. As in Huong's study, it was reported that the predictive power of language assessment was higher in the first semester than it was in the second semester; also the homogeneity of this sample was suggested to have positively affected the power of predictive validity.

In contrast to the above studies, Kerstjen & Nery (2000) found no significant correlation between the scores of a group of students in an Australian university (*n*=113) in IELTS and their GPA in the Business Faculty courses of the first semester. However, they did find a significant correlation between the writing and reading skills on one hand, and students' GPA on the other. They stated that "while the listening test was not significantly correlated to academic performance, students and staff … highlighted the importance of the listening skills in the first semester study" (p.85). The non-significant result was attributed to various factors; it was claimed that "sociocultural and psychological factors such as learning and educational styles, social and cultural adjustments, motivation and maturity, financial and family pressures have an influence

on the academic outcomes of international students in their first semester of study" (p. 85).

Breeze and Miller (2008), in a Spanish university, recruited 289 students from the faculties of Humanities, Law and Medicine and focused on the role of listening proficiency in predicting academic performance. The study used (1) students' self-reports of the difficulties faced in an academic environment, and (2) interviews with most of the students and some faculty members. The results showed weak positive correlations between the students' scores in the listening section of IELTS and GPAs in general $r = 0.34$, $p > 0.05$ in Humanities, $r = 0.28$ in Law, $p < 0.01$ and $r = 0.257$, $p < 0.05$ in Medicine. However, the correlations between the IELTS scores and students self-assessment ranged from moderate to strong, $r = 0.92$, $p < 0.01$, in Humanities, $r = 0.45$, $p < 0.01$, in Law, and $r = 0.34$, $p < 0.01$, in Medicine. Based on these results, it was recommended that the university should consider the students' evaluations of their coping abilities when looking into the entry requirements for English medium courses.

3.4.2. Studies on the Predictive Validity of TOEFL

This section presents some findings on the predictive validity of TOEFL. Like the previously discussed studies on IELTS, these reported varying levels of TOEFL predictive validity. Table 3.3 shows a sample of these studies.

Table 3.3. Some Studies on Predictive Validity of TOEFL

Study	Country	Number of Participants	Type of Correlation	Correlation
Vinke & Jochems (1991)	Netherlands	90 Indonesian students in Engineering	TOEFL with GPA	TOEFL<450 =0.09**
				TOEFL> 450=0.5**
Cho & Bridgeman (2012)	USA	2,594 graduate and undergraduate students	TOEFL and GPA	Graduate students=0.16*
				Undergraduates= 0.18*
Al-Musawi & Al-Ansari (1999)	Bahrain	86 undergraduate students, specialised in English language studies	TOFEL and GPA/ENGPA***	GPA=0.50**
				ENGPA=0.70**
Maleki & Zangani (2007)	Iran	Undergraduate students, specialised in English language studies	TOFEL and GPA	0.48*

*p<0.05
**p<0.01
*** Students' GPA in English Language Major

Vinke and Jochems (1991) correlated the scores of 90 Indonesian students in TOEFL with their GPA scores in Engineering courses at Delft University, the Netherlands; and found that the strength of the correlation coefficient depended on the range of TOEFL scores and age of participants; the TOEFL scores which were lower than 450 points showed a weak correlation coefficient (r= 0.09), but a strong coefficient (r=0.5) was found above 450 points. The correlation was higher for the participants who were younger than 33 years of age (r= 0.64, age <33, and r=0.38, age>33).

Cho & Bridgeman (2012) reported a large-scale study of 2,594 undergraduate and graduate students in ten universities in the US. The students' scores in TOEFL, iBT, GRE, and GMAT were correlated with their GPA by their disciplines and academic status (graduate/undergraduate). There was only a weak correlation (r= 0.16 for graduate students, and r= 0.18 for undergraduates) between language proficiency and academic achievement. Despite the small value of the correlation coefficients, the authors affirmed

that “even small correlations or seemingly trivial amounts of variance explained may be an indication of a meaningful relationship between two variables” (p. 439). They claimed that “their first year of university education were less likely to reflect academic performance within a single discipline” (p. 428) and recommended exploring this topic in longitudinal studies.

Al-Musawi & Al-Ansari (1999) compared the predictive validity of TOEFL to the predictive validity of the First Certificate of English FCE. The sample consisted of 86 undergraduate students from the first and second year of an English Language and Literature programme at the University of Bahrain. The students’ scores in each test were correlated with their GPAs using multivariate regression. The correlation coefficients of FCE with the GPA and ENGPA (GPA for English language courses only) were $r= 0.69$ and $r=0.84$, $p< 0.01$ respectively. The correlation coefficients for TOEFL and the GPA and ENGPA were $r=0.50$ and $r= 0.70$, $p< 0.01$. The authors concluded that the students’ achievement was more strongly correlated with their scores in FCE than in TOEFL (Al-Musawi & Al-Ansari, 1999, p. 397). However, this conclusion was later criticised for the analysis procedures implemented; Cho & Bridgeman (2012) stated that “their [Al-Musawi & Al-Ansari’s] interpretation should be taken in the light of the analytical approach used in the study. Because both the FCE and TOEFL are measures of English Proficiency, the sub-scores on the two tests would have been redundant creating a co-linearity problem in the analysis” (p. 423).

Another study that investigated the predictive validity of TOEFL using a sample of undergraduate students specialised in English Language and Literature was Maleki and Zangani’s Study (2007). They reported a similar correlation coefficient ($r=0.48$, $p <0.05$) to that reported by Al-Musawi & Al-Ansari. In general, studies that highlighted various facets of language testing predictive validity have indicated that the predictive validity has a weak correlation with general academic achievement but a strong correlation with second language specific academic achievement when the language of study is also the subject matter (Davies, 2008; Graham, 1987; Huong, 2001; Stansfield & Hewitt, 2005).

3.4.3. Studies on the Predictive Validity of In-House Language Tests

A number of studies have explored the predictive validity of language assessment instruments other than IELTS and TOEFL, and reported similar inconsistent results.

Table 3.4. Some Studies on Predictive Validity of In-house Language Tests

Study	Country	Number of Participants	Type of Correlation	Correlation
Davies (1990)	UK	310	ELTS, ELBA and EPTB with GPA	0.30**
Lynch (2000)	UK	475 international students	TEAM1 with GPA	0.32*
		291 international students	TEAM 2 with GPA	0.28*
Jochems, et al. (1996)	Netherland	170	Dutch exam and GPA	0.36**

*p<0.05
**p<0.01

For instance, Davies (1990) studied the predictive validity of three English language proficiency tests for students coming to study in UK universities. He "accumulated subject cases that is data from students taking the test and then collected also their academic grades and the proficiency judgments made on them later in their studies by their directors, advisors and supervisors" (p.47). Davies pointed out that generally, the predictive validity of most English language proficiency tests is low "with a correlation of about 0.3" (p.47). To understand this result, he listed some variables that might have influenced students' English language performance; one of which was individual learning of the language that took place between the administration of the language test and academic test. Therefore, he tested the student's English language skills again concurrently with their academic course tests to identify any improvement in the students' language skills levels. However, the results showed that there was "no evidence that causes us to claim that predictive validity is higher than 0.3 and we therefore need to explain why it is that language plays so small a part in academic study, only about a percent of the variance" (p.47).

Another example, this time with an in-house test, is Lynch's (2000) study which explored the predictive validity of the Test of English at Matriculation (TEAM) used at the University of Edinburgh. Students' scores in TEAM 1 (i.e., the 1989-92 cohorts: *n*= 475) were compared with different students' scores in TEAM 2 (i.e., the 1993-97 cohort: *n*=291), then each set was correlated with the same students' scores in their academic courses. The scores in the listening sections of both versions of TEAM correlated more with their scores in the academic courses than did their scores in the other TEAM sections (vocabulary and writing). The predictive validity value of TEAM1 (r= 0.32) was higher than that of TEAM 2 (r= 0.28). It was reported that the correlation value and significance considerably differed from one discipline to another: it was non-significant in the Faculties of Arts and Veterinary Medicine but significant in the Faculties of Law and Social Sciences (i.e. r=0.32 and 0.0.23 respectively). Lynch concluded that TEAM was effective for its purposes, which were to identify those who have achieved the minimum score for acceptance in academic faculties but whose language proficiency have not yet developed; and to provide data to decide on future language courses to assist those students in particular areas. This study, like others' referred to, identified differences in the strength of tests' predictive validity according to test takers specialisations and fields of study.

The predictive validity of assessment in languages other than English language has also been investigated. For example, Jochems et al. (1996) investigated the association between proficiency in Dutch as represented by students' scores in a Dutch language test and academic achievement as represented by passing an academic examination, together with the time taken to achieve this. They found that the success rate of non-Dutch speakers was similar to that of Dutch speakers, but the former group needed more time to pass the academic examination. The participants came from three faculties: Electrical Engineering, Computer Science, and Mechanical Engineering. The correlation coefficient between language proficiency and academic achievement was found to be r=

0.36, which is similar to the value of English language tests predictive validity reported in a number of studies presented above.

3.4.4. Methodological Limitations

In almost all language assessment predictive validity studies, two methodological aspects have been identified as having an effect on the findings. These are the range of levels of the participants and using GPA as a measure of academic achievement. Though it has been suggested that these two factors might have a negative influence on the validity of the findings, current studies are still influenced by these factors in ways which are explained below.

3.4.4.1 The Range of the Participants in Predictive Validity Studies

Graham (1987) suggests that the lack of consistency in the correlation coefficients between language proficiency and academic achievement could be a result of limiting the participants in predictive validity studies to those who have managed to pass a language test. She explains that when exploring the predictive validity by correlating the results of two assessment instruments, it has usually been the case that only the students who have passed the requirements of the assessment instrument in question have been allowed to undertake the other instrument and this seems to have resulted in depressing the predictive validity correlation coefficient. Graham used data from two studies in which the scores of the students with low language proficiency showed a strong correlation and the scores of those with high language proficiency showed a weak correlation. In line with this argument, several studies presented above note that the power of the language tests to predict academic achievement increases with lower scores. In general, this proposition is difficult to verify in higher education contexts where proficiency in English language is a gatekeeper and only those who reach a certain level are allowed to embark on academic study. The increased influence of language proficiency on academic achievement at the lower levels is still a matter of debate.

3.4.4.2. Using GPA as a Measure of Academic Achievement

Though GPA is widely used as a tool for measuring academic achievement there are some problems with this use. Graham comments that some researchers believe that GPA is not a valid indicator of academic achievement as it does not take into account the number of courses taken. Similarly, Jochems et al. argue that "a problem with the use of GPA is the fact that GPA's may be calculated over different periods of time" (Jochems *et al.* 1996, p. 326). Fox (2004) identifies other problematic issues with GPA as a measure of academic achievement, and asserts that it is not a static measure but a varying one that depends on various factors such as the students' social, financial, and academic circumstances; it was argued that students with more time and support for study were more likely to obtain a better GPA. Houng (2001) highlights other problematic issues in using GPA; he discusses the difficulties he has faced when trying to create one system to unify different types of GPAs used in the universities covered by his study. Cho and Bridgeman (2012) accept these problems of using GPA, and argue that it is the most suitable measure of academic achievement for this kind of study. They explain that other measures are problematic too, for example, using teacher evaluations as a measure of academic success could be criticised for its inherent subjectivity. In general, most of the studies on language assessment predictive validity recognise the pitfalls of using GPA as a measure of academic achievement, and argue that triangulating the findings from other sources or measures of academic achievement such as teachers' evaluations or students' self-evaluation reports along with the GPA would enhance the validity of the findings.

3.4.4.3. Non-linguistic Limitations

Several studies on the predictive validity of language assessment suggest that there are various non-linguistic factors that influence academic achievement and depress the correlation between language proficiency and academic achievement. The most common factors, which recur in several studies, are study skills, academic disciplines and

personal traits. Philips suggests that the difficulties students face in English-language tertiary education exist because they are "unfamiliar with the English sociolinguistic strategies of academic usage, and in fact attempted to use those of their former culture" (1987, p. 78). Woodrow's results substantiate this suggestion; she states that "students reported that finding resources is the biggest problem they faced" (2006, p. 62). Though Cotton and Conrow (1998) found that the writing assignments and reading academic texts were rated as the most difficult skills, they also reported that study skills were rated as moderate in difficulty. It seems that the students' study skills in tertiary education have a strong impact on their academic achievement and consequently on the strength of the correlation between language proficiency and academic achievement.

Furthermore, Graham (1987) reckons that there is a minimum level under which language plays a major role in academic achievement; she points out that this level could be different from one programme to another. Several studies have found that the association between language proficiency and academic achievement varies according to the discipline or academic programme in question. For instance, Jochems et al. state that "the relationship between foreign language proficiency and academic success - as expressed by means of a correlation - is higher in general for non-technological studies than for technological studies" (1996, p.326). Other similar studies that found that the predictive validity of language assessment varies from one discipline to the other were mentioned in sections 3.2 and 3.4.1. Moreover, personality traits are argued to have a similar impact on the strength of the association. Zabihi (2011) reports that "the results showed significant relationships between personality traits and proficiency as well as achievement scores".

As this study attempted to evaluate FP assessment in the Colleges of Applied Sciences, the presentation of the literature on validity theory, validation frameworks, programme evaluations and predictive validity has led the scope and methodology of this study. The questions of this study focused broadly on two areas: evaluation of assessment instruments and predictive validity. Validation studies usually focus on stated claims and uses of tests and work their ways back to practices and theories that support them (i.e.,

Bachman,2010; Kane, 2011), or study the different facets of validities including the consequences of tests in two phases (i.e., Weir, 2005). This study aimed at viewing the assessment instruments as a programme not only as tools. This view allowed including the opinions of stakeholders and studying the objectives, outcomes and needs of the programme. Certain validities such as construct validity and concurrent validity were not primary in this study; therefore, a full-fledged validation study was not used in this thesis. The type of evaluation that matches the focus of this study is impact evaluation which provides the researcher with a range of approaches. The predictive validity of FP assessment is an important part of this study as it informs on one of the main objectives of FP assessment: identifying able students to take on academic studies. Therefore a validation study was included in the evaluation of FP assessment.

Furthermore, it is crucial to consider the methodological and non-linguistic factors that affect the predictive power of language assessment in order to: first, design a methodology that takes into consideration these factors, and second, understand and attempt to explain the FP assessment predictive validity results reported in Chapter 10. The predictive power of FP assessment has serious implications with regards to the requirements of higher education. The main implication is increasing or decreasing the level of entry to academic study. Following Fox's argument (2004) that predictive validity studies are not "futile line of inquiry" and Davies's (1990, p.48) advice to explore why language has such a small part in predicting academic achievement, the present study attempts to explore one of the ways forward in this area of research by focusing not only on the strength of the correlation coefficient, but also on the factors that possibly have affected this correlation. These factors are: gender, specialization, and self-evaluation. It also investigates the linguistic needs of the First Year courses and assessment and compares it to what it is offered in the Foundation Programme.

3.5. Chapter Summary and Conclusion

In Sections 3.2, 3.3, and 3.4, the literature on two seemingly distinct fields of research (i.e. assessment validation and programme evaluation) was explored. The first section

compared the former view of validity and validation to the currently dominant one. It then described some frameworks of test validation and highlighted some issues concerning these frameworks. The second section presented purposes, types and paradigms of programme evaluation. From this discussion, common issues in both the fields of test validation and programme evaluation were identified, notably that of social consequences/impact. Three topics were discussed under this heading: washback, policy making and consequences, and stakeholders.

The literature discussed in the first part of the chapter informs the methodology used in this study. The current study could be encompassed under the wider umbrella of *impact* evaluation; it follows an eclectic approach that utilises elements of objectives-based, needs-based, goal-free, and process-outcome approaches to conform to its objects and foci. The multiplicity of the data sources and the intertwined areas of focus necessitated using elements from all of the four discussed approaches to impact evaluation and employing a mixed methods research design (Section 4.3 for a discussion on the methods used in this study). For instance, the objectives-based approach was used in investigating whether the stated FP learning outcomes corresponded with the textbooks learning outcomes and test specifications; the needs-based approach looked into the Foundation Programme's (FP) correspondence with the students' linguistic needs in the assessment of First Year (FY) and its effectiveness in identifying the students who were ready to embark on academic study; the goal-free evaluation allowed exploring the intended and unintended outcomes of the FP assessment and students experiences of FY academic study. Last but not least, the process-outcome approach was used to reflect on the implementation of language assessment in FP and FY, not by using observation which is the main data collection tool (Owen, 2007), but by alternative equally valid tools such as interviews, focus groups, students' scores records, and analysis of documents, tests and textbooks. This eclectic approach to impact evaluation and validation of FP assessment is inspired by Owen's declaration that "a good evaluation has a touch of artistry and creativity" (2007, p. 277).

This discussion of the literature on the predictive validity of language tests offers necessary background information on previous research in language testing predictive validity. It assists in explaining some of the findings of this study which looks into the predictive validity of the FP assessment in Oman. Some of the limitations reported by previous studies were addressed through triangulating tools of inquiry and considering some factors influencing the strength of predictive validity such as academic discipline, mode of assessment, gender or college.

As has been indicated earlier, the literature on validity, validation, evaluation and predictive validity informed the strategies used in the study. First of all, this study adopted the unified concept of validity (Messick, 1989) that viewed validity as a comprehensive evaluative judgment of test scores interpretations. Therefore, issues such as stakeholders and impact were focused on in this study. Second of all, it validated not only stated claims about test scores but also studied facets of test validity. The literature review revealed that validation studies were either restatements of the old trilogy of validity or were solely concerned with supporting or defying stated claims about test scores. Following recent calls to employ evaluation approaches in validation studies (Norris, 2009), this study integrated impact evaluation approaches with a predictive validity study to answer the study questions. Also, different sources of information were obtained such as: test scores, stakeholders' views, and documents on test construction and use. The methods and questions of the study are discussed in the following chapter.

Chapter 4: Research Design

4.1. Introduction

In this chapter, the research questions, methodology, methods, design, implementation, data-coding and analysis are discussed. It has been indicated in the first three chapters that this study follows a mixed-methods approach to explore the study questions. This chapter delineates more on the meaning of mixed-methods research and explicates the reasons for choosing this framework. After that, it presents the study questions and the methods; these methods are discussed in detail in terms of their structure, design, piloting and implementation. Then, a description of the data analysis stage is provided; this description includes sections addressing the various data types collected. The last part of this chapter clarifies aspects related to the quality of this study namely: validity/trustworthiness and research ethics.

4.2. Using Mixed-Methods Research

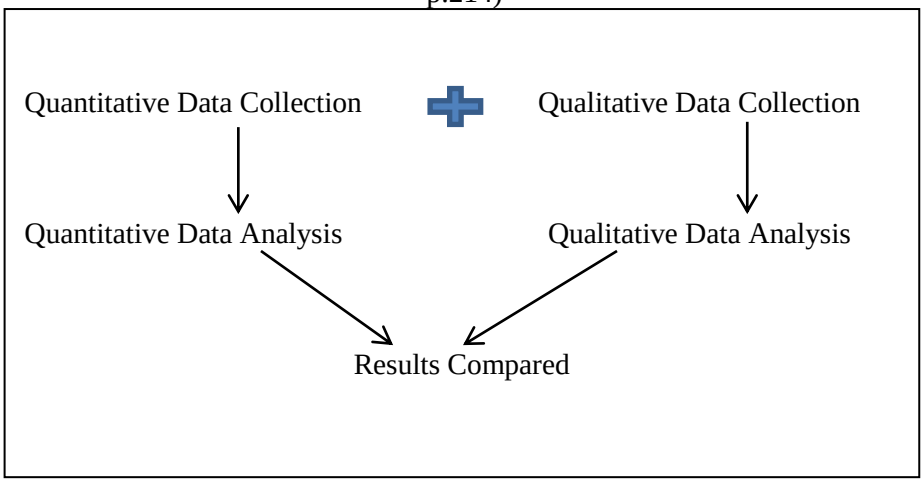
A mixed-methods study entails collecting and analysing quantitative and/or qualitative data using quantitative and/or qualitative methods (Dornyei, 2007). Considering that mixed-methods research uses both quantitative and qualitative methods to explore the questions of interest, researchers are urged to clearly state what and how the various methods are utilised and identify the links between the methods used and the topics of the study in the research design section.

> There is indeed a case for encouraging researchers to be explicit about the grounds on which multi-strategy research is conducted but to recognize that, at the same time, the outcomes may not be not predictable (Bryman, 2006a, p.110).

In this study, the nature of the questions necessitated using quantitative and qualitative methods; the effectiveness of the Foundation Programme (FP) assessment was explored by: firstly a quantitative evaluative study (Walliman, 2005) and secondly a correlational

study. The qualitative evaluation was used to explain and understand three main questions about: the structure and effectiveness of FP assessment, influences of policies and stakeholders on FP assessment, and stakeholder perceptions of the predictive validity of FP assessment. The correlational study investigated FP predictive validity and explored the factors that might have influenced its strength. Thus the research design utilised both quantitative and qualitative methods as is shown in Figure 4.1.

Figure 4.1 Concurrent Strategies in Mixed-Methods Approach taken from Creswell (2003, p.214)



Furthermore, one of the persuasive rationales for using a mixed-methods approach is that it allows triangulation of data from several methods. The aims of triangulation are to cross-check findings, and enhance the validity and reliability of a study through comparing the data collected from the different instruments (Bryman, 2006; Creswell, 2011; Creswell & Miller, 1997). In an advocacy of triangulation, Stufflebeam and Shinkfield argue that a case study design "addresses accuracy issues by employing and triangulating multiple perspectives, methods, and information sources." (2007, p. 183). Hammersley (2002) identifies two benefits of using a mixed-methods approach, or what he calls “methodological eclecticism”. These are: “facilitation” which denotes using one method as the basis of producing theories upon which the second method is designed, and “complementary” which entails obtaining different types of information that complement each other. Hammersley reports some researchers’ concern that the

“rapprochement” between quantitative and qualitative research types in a mixed-methods research might cause distortion of some of the theoretical features pertinent to each type. As a way out, he suggests downplaying the distinction between qualitative and quantitative research and focusing more on the strategies used in four “aspects”: “formulating the problems; selecting the cases; producing the data; and communicating the findings” (2002, p.173). This suggests that each aspect or stage in a research study could be individually characterised as qualitative or quantitative.

Furthermore, using mixed-methods research has been specifically called for in language assessment research (e.g., Brindley, 2003). He argued for using both qualitative and quantitative methods and sought to "end the false dualism between the two" (p.297). Thus, mixed-method research seemed to be the most suitable research design for this study considering the study questions, methods that could be used in this type of research, and endorsement of mixed-method research in the field of langauge assessment. The following section links the questions of the study to the methods used where the mixed-methods design becomes evident.

4.3 The Questions and Methods of the Study

It has been argued that epistemologically a mixed-methods approach is in harmony with a pragmatist stance where the research questions are given the prime focus in choosing data collection methods (Bryman, 2006b; Creswell, 2011; Onwuegbuzie & Leach, 2005). Tashakkori (2003) looked at how the concept “pragmatism” developed through time and stated that pragmatists “looked not at the origins of the idea but instead to its destination. What counted was not where you had been with an idea but rather where it took you”.

Considering this view, the study questions were deemed to be best responded to using both qualitative and quantitative data collection and data analysis. To clarify which methods will be used to investigate which questions, the questions are presented in the boxes below and ‘how’ and ‘when’ they were investigated is described beneath each box.

Box 4.1. Study Question 1

1. How well did the process of assessing students' English language performance, through classroom assessment and tests, function in the FP?
1.1. What processes and procedures were followed in writing and implementing the assessment instruments as depicted by the official documents?
1.2. How was the reliability and validity of FP assessment viewed by students and teachers?
1.3. How was the impact of the FP assessment perceived by students and teachers?
1.4. What were the differences between the 'continuous assessment' model used in the Academic English Skills course and the 'test' model used in the General English Skills course in terms of effectiveness, accuracy and preferences of teachers and students?
1.5. How did teachers perceive the centrally controlled assessment used in CAS?
1.6. What types (criterion/norm-referencing) of assessment were used? And how?
1.7. In all the above, were there any significant differences between the views of teachers and students with regard to the college, gender, self-evaluation and specialization groups?

The first main question and its following sub-questions were investigated in the first phase of the study using document analysis, teacher and student questionnaires, student focus groups, and teacher interviews. The questionnaires provided a general overview of teacher and student perceptions on these topics, whereas focus groups and interviews provided detailed accounts of their perceptions. To provide the context to understand and evaluate these perceptions, an analysis of some official documents on language assessment in the FP was conducted.

Box 4.2. Study Question 2

2. How did the assessment instruments correspond to the stakeholder wishes?
2.1. What were the national and international policies on teaching and assessing language that influence assessment in Oman?

2.2. What were teacher and student perceptions of the assessment tools' effectiveness and their roles in shaping language assessment?

Document analysis, interviews, and focus groups conducted in the first phase of the study were the main methods used to gather data to respond to the above questions. The stakeholder views were explored through interviews and focus groups, while policies were investigated through document analysis.

Box 4.3. Study Question 3

3. What was the predictive validity of the English language assessment for students' performance on the academic courses?

3.1. Did student performance in English language assessment in the FP correlate positively with their performance in academic courses?

3.2. Did the strength of the correlation between language proficiency and academic achievement differ when students' scores in English language tests only or continuous assessment only were used, instead of the accumulative scores in both?

3.3. Did the groupings by college, gender, self-evaluation and specializations show significant differences among the correlations between language proficiency and academic achievement?

3.4. How demanding were the learning outcomes and assessment of the academic courses in the FY on students' language skills?

Answering the above questions entailed a quantitative evaluation of English language assessment predictive validity; therefore, students' scores on the FP and their scores in the English language and academic courses in the First Year (FY) were gathered to conduct a correlational study. The students' scores were collected throughout two academic semesters. Also the students' scores in the last year of high school - prior to being admitted to the colleges - were collected to reach a comprehensive picture of their progress with respect to the English language.

Box. 4.4. Study Question 4

4. How did the stakeholders understand the relationship between the students' performance in the English language assessment and their performance in the academic courses assessment?
4.1. What were the teachers' and students' perceptions of issues related to the design, marking and impact of the English language assessment?
4.2. How did they think language accuracy should be considered in assessing academic assignments?
4.3. What were the teachers' and students' perceptions of the importance of the predictive validity?

As the questions above enquire about opinions on predictive validity and retrospective evaluation of the FP, questionnaires, focus groups and interviews were used in the second phase. The participants were mainly asked about the perceived difficulty/ease of FY study in terms of the language requirements and effectiveness of the FP as a pre-sessional programme to prepare them for the linguistic demands of FY courses.

4.4. Unexpected Events in the Locus of the Study

Before a further description of the methods used in this study, this section presents some challenges faced in the two locations of the study which affected the carrying out of the data collection. The first phase of this study was carried out in two (of six) Colleges of Applied Sciences (CAS), Sur College and Rustaq College, in the Spring 2010/11. In the same period, the Sultanate of Oman was going through a political uprising against issues such as corruption, unemployment, the low-quality and quantity of higher education, and high living expenses (see Section 1.3.1). Given that both students and their teachers mentioned these demonstrations quite often in the interviews, it seemed necessary to provide a concise description of how these uprisings affected the Colleges, and consequently this study.

The demonstrations took place mainly on the streets of some main towns and cities and, sometimes, they occurred in universities and colleges. In March 2011, when data was collected at Sur College, students started a demonstration that led to a strike; as a result,

the process of collecting data was stopped for three weeks and resumed when things were back to normal in the college. In the same month a student demonstration in Rustaq College resulted in a forced expulsion of the college dean by the students. The student demonstrations interrupted teaching and assessment schedules and caused some administrative and academic changes that were enforced by the Ministry. One of the changes that affected this study, as has been mentioned in Chapter 1, was allowing a group of suspended students to retake the FP two months before the final exams. This group consisted of students who had failed the FP assessment a year earlier and failed the entry exam administered at the beginning of every academic year, as well. The Ministry, in submission to the demands of the student demonstrations and in a political response to calm down angry students, granted them one more opportunity to pass the FP by permitting them to attend classes and undertake the FP assessment. Given that these groups were admitted very late to the FP and were very likely to fail, they were intentionally excluded from the main study. As expected, the FP results at the end of the semester showed that only a handful of these students successfully passed the FP assessment in each college.

4.5. The Methods of this Study

Document analysis; teacher questionnaires; student questionnaires; teacher interviews; and student focus groups were the five main data collection methods used in both phases of this study. How the methods were designed and implemented is discussed in the subsequent sections.

4.5.1. Document Analysis

In this study, document analysis was not used with the other methods as a means of triangulation only, it was also utilised to provide "background and context, additional questions, supplementary data, a means of tracking change, and development and verification of findings from other data sources" (Bowen, 2009, p.30). Various documents on language assessment from several sources were collected to (1) understand the context of English language assessment according to CAS policies as well as national policies, (2) highlight changes in English language assessment

procedures in CAS over the past two years and how that had been reflected in actual implementation, (3) compare how English language assessment on the FP was portrayed in official documents and compare that to stakeholders' perceptions. The purpose of gathering the documents varied based on how these documents were used in the study. Some shed light on the processes and procedures of assessment such as: course curricula; test papers; marking scales; and continuous assessment activities. Other documents stated general aims, objectives and assessment instruments of the FP, such as: *the Assessment Handbook*, *the Foundation Programme Booklet*, and *the General Foundation Programme Standards*. The rest of the documents were collected to obtain a better understanding of students' assessment results in English language courses over a period of two years. These documents included the students' scores in the last year of high school (retrieved in September 2010), their scores on the FP (retrieved in June 2011) and their scores in the FY (retrieved in February 2012).

Yin (2003), as reported by Bowen (2009), noted that sometimes documents might not be accessible or 'deliberately blocked' and considered this as a disadvantage of using document analysis. In this study, the process of obtaining the documents was not straightforward. The lists of the students' high school scores required time and effort to obtain as they were considered confidential and were kept in the national Higher Education Admission Centre. After a prolonged process of negotiation and explanation of the study content and aims with some of the Centre's personnel, the high school scores for the students who participated in the study were made available to the researcher. The results, however, were identified by the students' civil numbers, not their names or college numbers, both of which were available and easily accessible to the researcher. It was explained that this was done in conformity to the confidentiality of information policy followed by the Centre. Centre personnel refused to disclose the scores identified by the students' names or college numbers. As a consequence, the students' names together with their civil numbers were requested from the Colleges Registration Offices' in another long process to identify the students' high school scores.

In spite of the various difficulties faced in obtaining the students' scores, most of the other documents were easily accessible. The following table shows examples of some documents and their sources in both the first and the second phases of this study.

Table 4.1. Some Documents Collected in Phases 1and 2

The First Phase	
Document	**Source**
Mid-term, final tests and continuous assessment tasks	Coordinators and coordinators' website
Student scores in high school	Higher Education Admission Centre and CAS Registration departments
Foundation Handbook, Assessment Handbook, and other papers that included instructions for teachers on how to conduct continuous assessment	Assessment Coordinators' website
Academic English Skills course and General English Skills course textbooks	Heads of English Language Departments within CAS
Student scores in English language courses	The English Language Programme Director
Marking scales	Foundation Coordinators
General Foundation Programme Standards	Oman Academic Accreditation Authority website
College Regulations, Student Handbook and descriptive statistics about CAS students	Registration department within the Directorate General office of CAS
The Second Phase	
Document	**Source**
English language course specifications, textbooks and academic courses specifications	Programme Directors of each specialization
English language assessment tests and tasks, academic courses tests and tasks, and marking scales	Programme Directors of each specialization
Student scores on the English language course	English Language Programme Director
Student scores on the academic courses	Registration departments within CAS

4.5.2. Student and Teacher Questionnaires

4.5.2.1.Student Questionnaires in Phase 1 and Phase 2

In Phase 1, the student questionnaire was distributed to all FP level A[10] students in their classrooms after they were given a short verbal description of the study and informed

[10] Level A in FP is the highest level of three. Each level is taken over a period of about four months. This study focused on this level, because most of the students undertaking this level were expected to undertake

that participation was voluntary and about the confidentiality of the information gathered. The same information was distributed in a leaflet that included a summary of the study and an informed consent form (see appendix 4.1). The students were given some time to read the leaflet before deciding whether to respond to the questionnaire and participate in the focus group or not. Few students declined participation in the study. In Rustaq College, 127 students filled in the questionnaire out of 155 students enrolled on level A of the FP. In Sur College, 57 students out of 65 students enrolled in level A of the FP participated. These students were requested to hand the completed questionnaire to their teachers at the end of their classes.

The questionnaire consisted of six main topics, six sub-topics and 25 items. The development of the questionnaire went through several stages. The first stage involved writing the questions based on the topics of the study and previous research on the same topics. The main topics were decided upon first and then a number of items in each topic were written, having consulted related literature and the study's objectives to refine them. The items on the *Social Impact* sub-topic (see appendix 4.3) were guided by the findings on the impact of tests reported in Shohamy (2001), whilst the items on the *Political Impact* sub-topic were guided by Dale's study (1999). For instance, Shohamy reported that some social aspects of the impact of tests on students were feelings of unfairness, stress, fright or helplessness; she stated that:

> while test takers perceive tests as powerful, they see themselves as power-*less*, realizing that they have as little control over the requirements to take tests as over their consequences … This may explain why test takers often perceive performance on tests as 'pure luck', like a supernatural power they have no control over, with no understanding of the meaning of the results (p. 14).

These aspects were adapted and formed into questionnaire items to explore the social impact of the FP assessment as perceived by students. The items were then revised in terms of language and conformity to the study questions (see appendix 4.3).

---

academic courses the following semester which fits well with the purposes of this study that investigates the effectiveness of FP assessment in the transition period from FP to FY.

The questionnaire was designed to include three main sections, the first included questions about demographic data (e.g. college, gender, age, specializations); the second section included the items, and the third section included open ended questions - this section did not generate much response thus was not included in the data analysis. The questionnaire was developed first in English then translated into Arabic by a professional translator[11] to be administered to the students. The English language version was used to link the questionnaire items to the findings of previous studies on areas of interest, receive reviews on the plausibility of the content, and append it in this thesis for future reference. The Arabic version was used with students to eliminate any hindrance caused by the language of the questionnaire. Six months prior to conducting the study, the Arabic version was e-mailed to a sample of students in three colleges who were asked to respond to it and report any difficulties in understanding the questions. The comments received were incorporated to generate a pre-pilot version of the questionnaire (see appendix 4.7 for a sample of an English and Arabic versions of the questionnaire). The response rate to and results of the pilot questionnaires are discussed in Section 4.6.1.

In the second phase of the study, a similar procedure and structure was followed to produce a student questionnaire. The questionnaire included six topics and 21 items as shown in appendix 4.4. The focus of the questionnaire was on exploring student perceptions of the Foundation Programme English language assessment in retrospect, a semester after they had successfully passed the FP. Their evaluations of the English language assessment whilst on the FP and the adequacy of their language levels for their FY study were elicited in topics 1 and 2. Two aspects of assessment validity namely predictive validity and construct validity were the target of topics 3 and 4. The questionnaire items were adapted from the issues of predictive validity raised in academic sources such as: Baker (1989), Brown (1996), Davies (1995) and Fox (2004). The items were reformulated to make them fully applicable and understandable. The impact of the FP assessment was explored in topic 3; the items on social impact were

[11] The student questionnaires were translated by a professional translator to ensure that the language level used in the questionnaire is readable and comprehensible to the students before they were piloted. No translation services were used in this thesis elsewhere.

inspired by Shohamy (2001): *Power of Tests*. The items on political impact were adapted from Ball's (1998) views on international education policy. [Though the questionnaire divided validity into several topics, these topics are intended to be indicators of the overall assessment validity following Messick's unitary understanding of validity (see Section 3.2). Evaluation of validity should also take into consideration other pieces of evidence from stakeholders, such as teachers and students, using qualitative data generating methods, such as interviews and focus groups (Hamp-Lyons, 1997). This thought will be revisited in Chapter 11].

Another issue that the questionnaire focused on was student perceptions of how/if language aspects were assessed in the written assignments of the academic and English language courses. The items in this questionnaire were written based upon related literature. For instance, some of the items on the sixth topic, about assessing language aspects in academic courses, were adapted from a questionnaire used by Norton and Starfield (1997) conducted at the University of Witwatersrand, South Africa. The questionnaire was used to investigate "the extent to which proficiency in written English is perceived to be assessed in academic writing" (p.278). The adopted items were adjusted to suit the objectives, context and participants of the study.

4.5.2.2. Teacher Questionnaires in Phase 1 and Phase 2

The English language teachers in both Colleges were given a written description of the study and asked to hand back the informed consent that explained voluntary participation, anonymity and confidentiality of the participants' identities, when they agreed to participate. In the first phase, after 25 out of 34 teachers from both colleges agreed to participate, the questionnaires were distributed and teachers were contacted to arrange for semi-structured interviews.

The questionnaire used in the first phase consisted of six main topics and seven sub-topics as displayed in appendix 4.5. The topics of the questionnaire were selected to generate required information for the purposes of this study, and the items were based on related literature. The *Reliability* and *Validity* items were formed based on the

definitions of the two concepts available in pertinent literature (e.g., Brown, 1976; Baker, 1998; Davies, 1990) and accustomed to best suit the context of the study. For example, assessment tasks' reliability refers to the consistency of the tasks to assess the same skills among different groups, and the consistency in using specific criteria to evaluate performance of individuals (Alderson, Clapham, & Wall, 1995). Corresponding to this definition, the items in the *Reliability* topic were general statements about (1) the reliability of the rating scales, (2) the reliability of the assessment tasks, and (3) satisfaction with the reliability of the assessment instruments. The current literature on assessment validity tends to consider reliability, content validity, face validity, and construct validity and impact as facets that contribute to the understanding of a comprehensive concept of validity (e.g., Messick, 1989; Hughes, 2003; Weir, 2005). As Messick (1996) asserts, from the unitary perspective of validity, separate types of validation cannot be taken as sole measures of the overall validity of an assessment instrument. However, these facets are addressed separately in the questionnaire to facilitate understanding perceptions on each validity facet. The perceptions of the students and teachers should be understood as evidence on the face validity of FP assessment.

In the second phase, the teacher questionnaire included six topics, four sub-topics and 27 items (see appendix 4.6). The production of the questionnaire content went through several stages similar to those followed in producing the first phase questionnaire in terms of consulting the literature about the content and checking comprehensibility before conducting the pilot study. Some of the studies that were referred to when writing the items were: Davies (1995), Fox (2004) and Norton and Starfield (1997). The items on *Predictive Validity*, for instance, focused on teachers' perceptions of whether student language levels had an influence on their academic achievement. Though this topic has extensively been researched (e.g., Hill, Storch, & Brian, 1999; Huong, 2001; Xu, 1991), it was focused upon in this study to explore it from a different dimension (i.e., teacher and student perceptions) in addition to the other common dimensions (i.e., student scores, and GPA).

4.5.2.3. Teacher Semi-Structured Interviews in Phase 1 and Phase 2

In selecting the participants for the teacher interviews in the first and second phases, stratified sampling was used and followed by convenience sampling that was subject to teachers' availability and willingness to participate (Bryman, 2004, p. 334). In the first phase, only English language teachers who taught on the FP were invited. From both colleges, 30 FP teachers were asked to participate. In the second phase, 34 teachers from the English language, Information Technology (IT), International Business Administration (IBA), Communication Studies (CS) departments were invited to participate in the study.

In the first phase, when English language teachers completed the questionnaire, they were asked to participate in semi-structured interviews. Only 19 teachers agreed to participate from both colleges in the first phase and 23 teachers participated in the second phase. Each teacher was scheduled for a 30-40 minute interview subject to their timetable, availability and approval. In each interview, teachers were assured of the confidentiality of the information yet again and were asked permission to record the interviews. A list of four main questions was used to guide the interviews (see Box 4.5), however, sometimes the order of the questions was changed and probes were used as appropriate**.**

Though the interviews were conducted in the second semester of the academic year 2011, four of the interviewed teachers were new to their college due to the high rate of teacher turnover within CAS. Two teachers from Sur College were interviewed but were not able to respond to the questions at that stage and asked to be interviewed again at a later date. They explained that they were new and were not yet aware of the assessment of their courses. In the second visit to the college, two weeks from the final test, one of the two teachers had already left the college whilst the second one was available for a second interview. The foundation coordinator in Sur College was also interviewed again to attain a more comprehensive view of his impression of the mid-term test and continuous assessment, two weeks before the final test. In Rustaq College, most teachers were willing to complete the questionnaire; however, they were reluctant to participate

in the interviews. Only 9 out of 29 teachers agreed to participate. The researcher was asked to leave teachers' offices twice and some appointments to conduct interviews were broken. Regardless of these difficulties, 19 willing teachers from both Colleges participated and shared their perceptions in the interviews.

Box 4.5 displays the main questions that all participants were asked in the first semester. Sometimes, different prompts were used with different participants depending on the progress of the interview. The topics of the questions correspond to those used in the questionnaire and required detailed responses. The interviews were used to provide and add elaboration, explanation or context to the teachers' responses in the questionnaire.

Box 4.5. Main Questions of Teachers' Interview (Phase 1)

1. How do you assess your students' performance in English language?
2. What is your role in writing the assessment instruments?
3. What do you think is the best way to assess students' language competence?
4. What do you think of the General Foundation Programme standards (GFPs) implemented in term of language assessment?
5. How do you see the integration of tests, quizzes and projects as a unified means of assessment?
6. How does your role impact on the tests? When do you think you are listened to in regard to test design and writing? Why? How should it be?

In the second phase of the study, not only English language teachers were interviewed but also teachers who taught in other academic departments namely Information Technology, International Business Administration, and Communication Studies. The interview questions for the English language teachers were very similar to the interview questions for the academic teachers, except that the second question in the English language teachers' interview was not included in the academic teachers' interview (see box 4.6 and box 4.7). The same questions were used in both interviews to compare and contrast the teachers' views and investigate any differences in the views that could be linked to the different disciplines.

The topics that were raised by the interview questions were about teachers' opinions on the (1) predictive validity of language assessment, (2) language related difficulties faced

in the FY, (3) marking language in written assignments, and (4) the impact of language assessment. As was the case with the items in the questionnaire, the interview questions were driven by the focus of the study and were informed by related literature. They were also revised for comprehensibility and appropriateness before being piloted.

At the beginning of each interview, the teachers were informed about the anonymity, confidentiality and voluntary nature of their participation in the interviews. In most cases, the questions were followed in the order presented in the boxes below, but in some cases the order of the questions differed to go with the flow of the interview and the participants' responses. For example, in responding to a certain question, if a participant discussed something on a different but related topic, the interviewer went along with the participant but returned to the first topic when the second one was fully or appropriately discussed. Given the limited time and structured nature of the interviews, any deviation from the focus of the questions was minimised. This is also in line with the pragmatic view of this study where the study questions are the driving force in shaping the research design and implementation.

Box 4.6. Main Questions of English language Teacher Interview (Phase 2)

1. From you experience with your students inside and/or outside the classroom, how do you think students' English language levels affect their performance in academic courses in terms of:
   - different language skills;
   - their readiness to deal with FY academic courses
2. How do you ensure that the ways you use to evaluate students helps them in how they are evaluated in the academic courses?
3. Do you think that some language skills are more important than others for students' academic achievement? How and what skills?
4. How much do you consider content knowledge when marking written assignments?
5. Given that most FY courses are taught in English, I would like to know about the negative and positive aspects for students studying their specialization in English in terms of:
   - the international nature;
   - promoting Higher Education;
   - English as a gatekeeper.

Box 4.7. Main Questions of Academic Teacher Interview (Phase 2)

1. From you experience with your students inside and/or outside the classroom, how do you think students' English language levels help them in their academic courses in terms of:
   - different language skills;
   - their readiness to deal with FY academic courses.
2. How much do you consider English language skills (grammar, vocabulary…etc.) when marking written assignments?
3. Given that most courses in the FY are taught in English, I would like to know about the negative and positive aspects for students studying their specialization in English in terms of:
   - the international nature;
   - promoting Higher Education;
   - English as a gatekeeper.

4.5.2.4. Focus Groups in Phase 1 and Phase 2

> The focus group offers the researcher the opportunity to study the ways in which individuals collectively make sense of a phenomenon and construct meaning around it. (Bryman, 2004, p.348)

In the first phase, arranging to conduct focus groups was difficult and demanding due to concurrent student demonstrations and their hectic schedules. The students had 26 hours a week of classes; this was about five hours a day on average. Also the fact that male and female students preferred to attend gender specific focus groups as the pilot study revealed made it even more difficult to allocate two suitable timings for each of the groups. In an attempt to increase the number of students participating in the focus groups, and to avoid selective participation, all of the students who completed the questionnaire were invited to participate. Contrary to researcher expectations, a large number of students participated in the focus groups. In the first phase focus groups, 106 students participated and most of them were active in group discussions. Very few students did not participate at all in the discussions.

All focus groups were conducted in meeting rooms to avoid classroom settings which might have imposed teacher-student constraints. The participants were briefed on how focus group should function and about the roles of the facilitator and the participants. In the first focus group, the researcher recruited a student to be the facilitator but it did not work well as he lost track of the questions on several occasions and did not allow for pauses in discussions, so the researcher took the responsibility of the facilitation role. The questions and discussions were delivered in Arabic to eliminate any language barrier that could have inhibited students from freely expressing their thoughts. The discussions were videotaped to capture a better view of the live communication and body language and to be able to identify participants in the transcription stage later. In most groups, many of the participants directed their answers to the facilitator who intentionally avoided eye-contact with the person who was speaking. Generally, avoiding eye-contact compelled the speaker to look at other participants and resume the discussion with them. The students were informed that, in a focus group:

- the researcher should facilitate the discussion from time-to-time and should not dominate or tell the participants what to say;
- all participants should express their views at all times or at least express their agreement or disagreement about the issues discussed;
- allowing certain participants to dominate the discussion should be avoided as much as possible;
- there were no wrong or right answers, only their opinions.

The focus group discussions were guided by a small number of questions followed by probes whenever needed (see Box 4.8). When the discussions stopped or drifted away from the topic for two to three minutes, the facilitator used probes to return to the main discussion. The order of the questions used in the focus groups and the probes changed depending on the progress and flow of the discussions.

Box 4.8. Main Questions for Focus Groups Phase 1

1. How useful are the assessment instruments (classroom assessment and tests) in assessing your language levels? Do they reflect an appropriate image of your language level?

2. Why are you assessed via multiple instruments (classroom assessment and tests)? How do they work together? And how do they help you?

3. How important is it for you to pass the FP? Why?

4. How does your voice impact on the tests you have? When do you think you are listened to in regard to tests? Why?

The following table specifies the number of participants, college, gender and length of the focus groups in Phase 1. Though it is recommended that focus groups are conducted for about an hour (Iowa State University, 2004), most of the focus groups in this study lasted for about 30 minutes on average because the number of questions was small and the students were not willing to stay longer due to their hectic schedules.

Table 4.2. Focus Groups in Phase 1

Group	College	Gender	Student numbers	Length/minutes
Group 1	Rustaq	F	12	53 min.
Group 2	Rustaq	F	8	32 min.
Group 3	Rustaq	F	16	32 min.
Group 4	Rustaq	F	9	35min.
Group 5	Sur	M	9	33 min.
Group 6	Rustaq	F	6	32 min.
Group 7	Sur	F	12	38 min.
Group 8	Rustaq	F	13	26 min
Group 9	Sur	M	8	38 min
Group 10	Sur	M	3	14 min.
Group 11	Sur	M	7	34 min.
Group 12	Rustaq	M	13	51 min.
Total			**106**	**418 min.**

In the second phase of this study, the 181 students, who had participated in the first phase and had successfully passed the FP, were given the questionnaire and asked to

participate in focus group discussions. Though all of the questionnaires were returned, only 80 students participated in focus groups, this number is less than the number of students who had participated in focus groups in Phase 1. The students explained their reluctance to attend saying that they were inundated with quizzes and assignments towards the end of the academic semester. Focus groups were conducted in the last few weeks of the autumn semester in 2011. This time was chosen to allow students enough time to experience assessment instruments used in the courses and be informed about the courses in the FY and to self-evaluate the appropriateness of their English language levels for the study in the FY. To overcome this hurdle, some teachers were contacted about the possibility of conducting focus groups in the second hour of their two hour lectures explaining that few students showed up because they were swamped with the end of the semester quizzes, reports and presentations. Several teachers welcomed the idea and encouraged their students to attend the focus group sessions. Regardless of the teachers' encouragement, some focus groups included only a very few students and sometimes the students did not show up at all for scheduled and agreed appointments. Bryman (2004) warned of the 'no shows' in focus groups and recommended to continuously over-recruit. Following this advice, 15 focus groups were conducted and videotaped in a similar manner to the ones conducted in the first phase (see Table 4.3). The questions used in Phase 2 focus groups are displayed in Box 4.9.

Table 4.3. Focus Groups in Phase 2

Number	College	Gender	Number of Students	Specialization
Group 1	Sur	F	4	Communications
Group 2	Sur	F	3	Communications
Group 3	Sur	M	3	IT
Group 4	Sur	M	4	Communications
Group 5	Sur	M	6	Communications
Group 6	Sur	M	3	Communications
Group 7	Sur	M	9	Communications
Group 8	Sur	M	3	IT
Group 9	Rustaq	F	3	IT
Group 10	Rustaq	M	9	Business & IT
Group 11	Rustaq	F	4	Business
Group 12	Rustaq	M	7	English language
Group 13	Rustaq	M	13	Business
Group 14	Rustaq	F	5	English language
Group 15	Rustaq	M	5	IT
Total			**81**	

Box 4.9. Main Questions Used in Focus Groups Phase2

1. What were your expectations of the FY in terms of the language demands?
    - How did you find it?
    - Were your expectations affected by the fact that you passed the FP?
    - Are your language grades and academic performance similar?

2. How do you think your language level is affecting your performance/grades/results in academic courses if at all?
    - How should it be?
    - Do you feel that this is fair or reasonable?

3. How you think lecturers in the academic courses penalise or reward you for your written language used in (activities/projects/exams)?
    - How should it be?
    - Is it fair?

4. Do you think that teachers should penalise or reward you for your written language?

4.6. The Pilot Study for the Methods Used in Phases 1 and 2

The pilot study took place in October 2010, while, the questions used in the focus groups and interviews were piloted in February 2011. The questionnaire pilot was earlier than the focus groups and interview pilots to allow for statistically analyzing the results and testing inter-item reliability of the questionnaire. Although the main study was conducted in two Colleges, Rustaq and Sur, the pilot study took place in three Colleges namely Rustaq, Ibri and Nizwa. This was due to the difficulty in recruiting participants when they understood that their participation was for a pilot study not a main one as the course coordinators who participated in distributing the questionnaire explained. However, piloting the focus group and interview questions was less complicated and more straightforward; it was conducted in Sur and Rustaq Colleges a month before the main study took place.

4.6.1. Student and Teacher Samples in the Pilot Study in Phase 1 and Phase 2

In piloting the first phase questionnaire, 46 students and 31 teachers participated while 41 students and 15 teachers participated in piloting the second phase questionnaire. The questionnaires were e-mailed to course coordinators on the FP and in the FY who distributed them randomly to students and teachers. The coordinators collected the completed questionnaires and mailed them back to the researcher. Comments on the structure and clarity of the questionnaire from the students and teachers were e-mailed to the researcher, as well. The four tables below display general background information about the students and teachers who participated in the pilot stage.

The interviews and focus group questions in the first phase were piloted with 10 students and two teachers, whereas, the interviews and focus groups questions of the second phase were piloted with eight students and three teachers.

Table 4.4. Students' Sample in Piloting Phase 1 Questionnaire

Category	Sub-Category	Frequency	Category	Sub-Category	Frequency
College	Nizwa	16	**Age**	17 years	5
	Rustaq	14		18 years	33
	Ibri	16		19 years	7
				20 years	1
	Total	**46**		**Total**	**46**
Gender	Male	9	**Specialization**	IT	26
	Female	37		Design	6
				Communication Studies	14
	Total	**46**		**Total**	**46**

Table 4.5. Students' Sample in Piloting Phase 2 Questionnaire

Category	Sub-Category	Frequency	Category	Sub-Category	Frequency
College	Nizwa	12	**Age**	18 years	1
	Rustaq	15		19 years	28
	Ibri	15		20 years	12
				21 years	1
	Total	**42**		**Total**	**42**
Gender	Male	14	**Specialization**	IT	29
	Female	28		Design	5
				Communication	8
	Total	**42**		**Total**	**42**

Table 4.6. Teachers' Sample in Piloting Phase 1 Questionnaire

Category	Sub-Category	Frequency	Category	Sub-Category	Frequency
College	Nizwa	8	**Age**		
	Rustaq	12		20-30	6
	Ibri	11		31-40	7
	Total	**31**		41-50	9
Gender	Male	17		51-60	7
	Female	14		61+	2
	Total	**31**		**Total**	**31**
Nation-ality	Omani	4			
	Non-Omani	27			
	Total	**31**			

Table 4. 7. Teachers' Sample in Piloting Phase 2 Questionnaire

Category	Sub-Category	Frequency	Category	Sub-Category	Frequency
College	Nizwa	12	**Department**	Communi-cation	2
	Rustaq	3		IT	4
	Total	**15**		English	9
Gender	Male	11	**Nationality**	Omani	4
	Female	4		Non-Omani	11
	Total	**15**		**Total**	**15**

4.6.2 Piloting Student and Teacher Questionnaires in Phases 1 and 2

4.6.2.1. Student Questionnaires

Piloting the questionnaires included two stages, firstly analyzing the comments given by the students about the structure and comprehensibility of the questionnaire items, and secondly analyzing the results of the questionnaires for inter-item reliability of the questions in each category. The results from student comments for both phases showed that they seemed to have difficulty in understanding some questions which were later changed. One of the issues believed to cause difficulty related to translation as the questionnaire was administered in Arabic. In addition, their responses suggested that there should be changes in the first section about the age and specialization categories that were amended to include more sub-categories than had been given in the previous version. For instance, in the first section, the students were given four different ages to choose from ranging from 18 years old to 21 years old; however, some students commented that they were 17 years or 22 years old. So the age range was amended to 17 to 22 years old.

In the second stage, an inter-item correlation was administered to check the reliability of the questionnaire items. Pallant (2010) stated that it is normal to find low Cronbach alpha values in short scales of less than 10 questions. She also recommended reporting the inter-item values when the Cronbach's Alpha is found to be of low value. As each of the categories in both questionnaires consisted of less than 10 items, the Cronbach's

alpha and inter-item correlations were sometimes found to be low as shown in tables 4.8 and 4.9 The tables display the alpha values, inter-item correlation value and any subsequent changes made to the questionnaires that resulted from student comments, responses to the questionnaire items or/and reliability analysis of the questionnaire items. Splitting some questionnaire items into two is one example of the changes that were based on students' comments. For instance, some students were confused by certain items that did not address continuous assessment and tests separately; to overcome this confusion, the items were split to represent continuous assessment and tests in two different statements.

It is important to note here that though some of the items were re-statements of one or similar concepts, other items were about different but related aspects of a concept. In the first type where the items are reiterations, Cronbach alpha is expected to be higher than that of the second type. However, sometimes alpha was found to be similar and the inter-item range of the first type was found to be higher than that of the second type. For example, in Table 4.9, the items comprising the *Reliability* topic scored an alpha value of 0.66 similar to that of the *Social Impact* topic. However, the inter-item correlation range shows that the *Reliability* items (i.e., 0.38-0.49) were more consistent with each other than the Social Impact items (i.e., 0.09-0.77). This confirms the point that it is important to obtain a high alpha value when the items of a topic are re-statements of one concept, but it is expected and acceptable to obtain a low alpha value when the items address different aspects of a concept.

Table 4.8. Inter-Item Correlation for Student Questionnaire in Phase 1

	Topic	Number of Items	Alpha	Inter-item Correlation Range	Changes Implemented
Reliability		2	0.66	0.38-0.49	No change
Validity	Content	4	0.55	0.07 -0.5	The first question was divided into two that addressed tests and continuous assessment separately.
Construct	General	1	-	-	-
	Test	2	0.71	0.57	No changes
	CA	2	0.68	0.52	No changes
Preference for tests		2	0.48	0.32	No changes
Preference for continuous assessment		2	0.50	0.33	No changes
Satisfaction with current assessment tools		3	0.63	0.32-0.43	No changes
Impact	Social	5	0.66	0.09-0.77	The number of items was reduced to get better inter-item consistency since there are many items on this topic. Originally there were seven items. Item10 was divided into two items that addressed continuous assessment and tests separately.
	Political	2	0.71	0.55	Reduced to two from three for better inter-item correlation.

The table above shows that alpha values of internal consistency in each topic ranged from 0.48 to 0.7 and the inter-item reliability ranged from 0.07 to 0.77 showing moderate to strong internal consistency. Briggs and Cheek (1986) as reported by Pallant (2007) recommended the range of inter-item reliability to be of 0.2 to 0.4. Though the lower end of the inter-item reliability range was lower than recommended, most of the inter-item reliability ranges for the comprised topics fell within or higher than the

recommended range. Another thing that could be noted from the table is that the number of questions designated to *Validity* and *Social Consequence* were more than the questions designated to any of the other topics; this was because the questions dealt with different validity aspects. For example, in the *Validity* topic the questions were about four aspects of validity namely: face validity, content validity, construct validity and predictive validity. Each of the aspects included items that might shed some light on how students viewed the validity of the FP assessment instruments used, these views contribute to understanding the face validity of the assessment.

The table below displays the alpha values and inter-item reliability ranges of the questionnaire topics in Phase 2. It shows that the alpha values ranged from 0.18 to 0.56 indicating low to strong correlations between the items of each topic. Most of the inter-item reliability ranges fell within the recommended range but for two. The alpha values of the constituting items of topic number five was not calculated as each item focused on a slightly different idea from the other, though all of the items were about evaluating English language in the academic courses. The table also notes and explains any changes that occurred post-piloting the questionnaire.

Table 4.9. Inter-Item Correlations for Student Questionnaire in Phase2

Topic	**Number of Items**	**Alpha**	**Inter-item Correlation Range**	**Changes Implemented**
Construct Validity	2	0.37	0.23	One item was deleted
Dissatisfaction with FP assessment	3	0.23	0.18	One item was deleted and another was divided into two, one about the FP and one about the FY.
Adequacy of English language level	4	0.7	0.05-0.56	No changes
Consequence and Impact	3	0.45	0.26-45	No changes
Assessing English language and ideas in subject and English courses	7	-	-	Each of the items dealt with a different aspect of the topic, so inter-item reliability cannot be applied here.
Predictive validity	2	0.55	.38	No changes

4.6.2.2. Teacher Questionnaires

Unlike student questionnaires, the teacher questionnaires' wording and format were reported to be comprehensible by the teachers in the pilot study. Therefore, no changes were made to the wording or format of the questionnaire. The following tables shows the changes resulting from the inter-item reliability test applied to the Phase 1 and Phase 2 teacher questionnaires. In several instances in the two questionnaires, it was possible to omit an item in certain topics if it lowered the alpha value when there were other similar items that addressed the same area of interest.

Table 4.10. Inter-Item Correlation for Teachers' Questionnaire in Phase 1

Topic		**Number of Questions**	**Alpha**	**Inter-item Correlation Range**	**Changes Implemented**
Reliability		3	0.6	0.22-0.55	No changes
Validity	Content	3	0.8	0.45-0.74	One item deleted for being off topic
	Predictive	2	0.77	0.62	One item deleted for better correlation
	Face	3	0.52	0.14-0.42	One item was deleted for better correlation
	Construct	3	0.72	0.30-0.61	No changes
Tests are more reliable and valid		2	0.92	0.85	No change
Preference of centrality in writing assessment		3	0.71	0.33-0.64	One item was deleted for better correlation
Experience in marking and writing assessment		3	0.71	0.47-0.56	No changes
Impact	Social	5	0.78	0.13-0.66	Two items were deleted for better conformity
	Political	3	0.5	0.15-0.36	

Table 4.11. Inter Item Correlation for Teachers' Questionnaire in Phase2

Topic		**Number of Questions**	**Alpha**	**Inter-item Correlation Range**	**Changes implemented**
Reliability		4	0.72	0.55	One item was divided into two
Validity	predictive	5	0.59	0.1-0.5	No changes
	construct	2	0.5	0.32	One item deleted
Satisfaction with assessment	FP	3	0.48	0.31	One deleted
	FY	2	0.41	0.26	No changes
Assessing English in academic courses		5	0.79	0.24-0.78	No changes
		3	0.56	0.04-0.53	No changes
		3	0.61	0.014-0.51	No changes

We can see from the tables above that the alpha values and inter-item reliability ranges for Phase 1 are better than those for Phase 2. The first phase questionnaire shows high conformity rate, while the second phase questionnaire shows a lower conformity rate. This could possibly be attributed to the fact that in Phase 2 both the English language courses and academic courses teachers participated while only the English language courses teachers participated in Phase 1 in February 2011.

4.6.2.3. Piloting Student Focus Group Questions

The focus group list of questions was discussed with the supervisors of this thesis for appropriateness and with a translator for correctness before being piloted. One month prior to conducting the main study, two focus groups, consisting of six male students and four female students, were asked to participate in piloting the focus group questions. Originally, the researcher asked them to participate in mixed gender focus groups, but they hesitated and seemed to prefer gender specific focus groups, therefore separate sessions were arranged for the male and female students. The discussions took place in a classroom and were recorded. The main changes to conducting the focus groups that resulted from the pilot study were (1) videotaping the discussions instead of tape recording them, and (2) arranging for the focus groups to be held in a meeting room rather than a classroom. In the pilot focus group, it was difficult to identify which of the

students was talking from the tape recording, but when they were videotaped, the process was much easier. Conducting the focus groups in classrooms negatively affected the discussions when they were sometimes interrupted by passers-by or students who mistakenly entered the room; it also seemed to dictate a teacher student interaction instead of peer discussions. Participants were looking at, and communicating with, the facilitator instead of fellow participants. This was, to a certain extent, decreased when the focus groups were held in a meeting room where students were seated around an oval shaped table.

In piloting the questions for the second phase, four male students and four female students were invited to participate in two focus groups. The students were from different departments. In both groups, the students thought that the questions were clear. The female students suggested adding questions to the mathematics course assessment instruments which they were not happy about. These suggested questions were not integrated in the list of the questions used in the main study as they were deemed to be irrelevant to the focus of the study.

4.6.2.4. Piloting the teacher semi-structured interview

The semi-structured interview of the first phase was conducted with two English language teachers in the pilot study to evaluate how well the questions stimulated responses on the target area. In general, the questions were clear and probed teachers' experiences. However, the teachers' responses tended to be about only one type of assessment instrument, either tests or continuous assessment, instead of both. Later, it appeared that most teachers were teaching only one FP course, the General English Skills course, that used a mid-term and a final as its assessment instruments, or the Academic English Skills course, that used continuous assessment as its assessment instrument. Therefore, it became worth asking the teachers about the courses they taught before proceeding with the interview questions.

4.7. Data Analysis

> It is critical to advance the rationale of conducting mixed-method study … it is especially important to identify how the results will be integrated (or kept separate) in the research findings (Creswell & Miller , 1997, p.46).

Considering that this study is a mixed-methods study that follows a pragmatic stance in research, and considering that this study included correlating and evaluating elements, qualitative and quantitative analysis approaches were implemented. Thematic content analysis guided the process of analyzing the official documents, interviews and focus groups, whereas, descriptive and inferential statistics were used to analyze the questionnaires and student scores. All these approaches and methods of analysis are delineated below in three main sections: Thematic document analysis, thematic content analysis of the interviews and focus groups, and descriptive and inferential analysis of the questionnaires and student scores.

4.7.1. Document Analysis

The approach to document analysis was thematic analysis that is 'a form of pattern recognition' (Bowen, 2009, p32). Although in the design of this study a critical hermeneutics approach was intended to guide the document analysis, it was found to be impractical for the purposes of the study and types of documents collected. Critical hermeneutics as developed by Philips and Brown (1993) and Forster (1994) focused on both the context of the documents within which they were produced and the point of view of the author in generating common themes. Linking the themes to the context and authors' views was not chosen in this study for two reasons. First, the document analysis was one of four sources of data in this study; therefore, it was felt that applying similar codes to those generated by the interviews and focus groups would facilitate integrating data (Bowen, 2009). This does not mean that the codes used in the interviews were solely implemented in the document analysis. In fact, the interview codes were not imposed on document analysis but rather they were used to lead it, and other new codes

were created in the process of document analysis. Second, the author's views and context of the documents could not be identified for all the collected documents (e.g., student marks, and task specifications). Therefore thematic analysis was employed in document analysis to facilitate comparing and contrasting the results from different data sources. This comparison is intended to reveal the reality of what is presented in the documents. Atkinson and Coffey (2004) argued that documents are written with hidden purposes in mind and they could suppress some realities if they were to be displayed in public, so the writers warned that

> we cannot … learn through written records alone how an organization actually operates day by day. Equally, we cannot treat records - however "official"- as firm evidence of what they report (Atkinson & Coffey, 2004, p.58).

About 85 documents ranging from one page to fifty pages long were coded and analysed for common themes related to the study. To ease retrieving coded extracts from this large number of documents, Atlas ti. (i.e., a qualitative data analysis tool, see Figure 4.1) was used. The documents were divided into documents about Phase 1 and documents about Phase 2 and then they were uploaded into the software which was strictly used only to organise the documents and codes for faster retrieval. The analysis process went through several steps to generate themes that embodied the main issues on the quality of assessment writing and implementation in the FP. These steps are described below.

a. Initial reading and highlighting of possible important points.

b. Secondary reading that included forming a list of codes that either emerged while reading or were used in the interviews and focus group analyses.

c. Refining the codes by excluding the less common ones and the ones that were irrelevant to the subject of the study.

d. Uploading the codes to Atlas ti. The figure below shows a document in the coding process. The codes are on the right hand side and the document is on the left hand side. When a code is selected the linked extracts become highlighted.

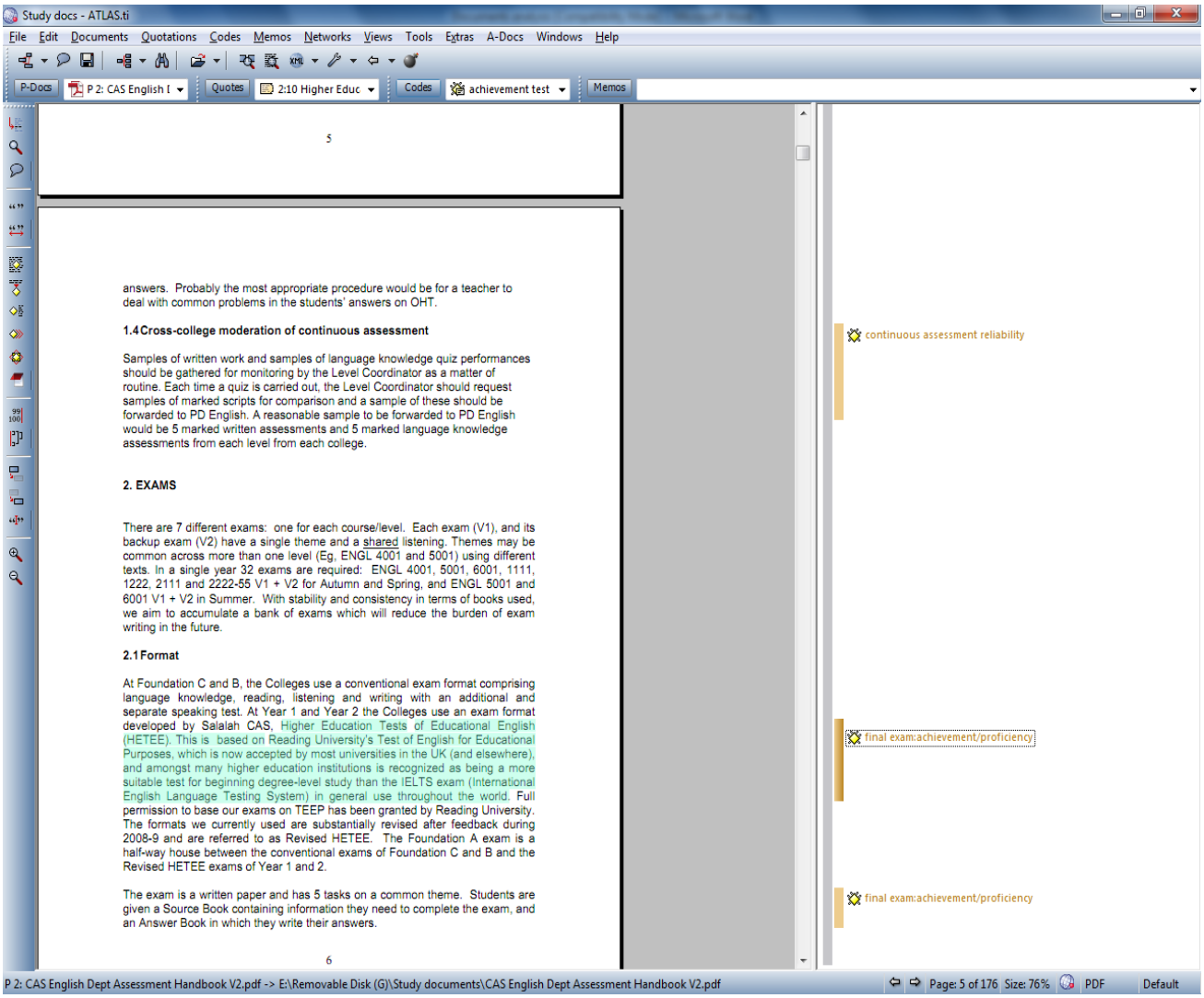
e. Reading the documents again prior to assigning the selected codes.

f. Coding the documents. Returning to the questions of the study to focus the codes.

g. Reading the extracts and organizing them into themes. Going back to the original texts to check if themes are appropriate and comparing them to the themes generated by the other methods to ensure that similar themes were focused upon in the analysis.

h. Writing up the results based on the themes found.

Figure 4.2. Assigning Codes to Texts in Atlas ti



4.7.2. Thematic Content Analysis of the Interviews and Focus Groups in both Phases

Though content analysis is sometimes linked to quantifying the elements of the content according to a set of categories in a systematic manner (Bryman, 2008), thematic content analysis is linked to qualitative data analysis. According to Bryman interpretation of Althiede's analysis (1996), ethnographic content analysis involved an element of "constant discovery and constant comparison of relevant situations, settings, styles, images, meanings, and nuances" (p.393). In this study, thematic (ethnographic) content analysis which focuses on "what is said rather than on how it is said" (Bryman, 2008, p.412) was used to analyze the transcripts of the teacher and student interviews.

Although the transcripts produced by the interviews and focus groups were all analyzed following similar parameters of thematic analysis, they were approached differently. The teacher semi-structured interviews were coded first and then the researcher transcribed the parts needed, while focus groups were all transcribed first and then coded. In the first phase, the interviews were coded and analyzed in their audio form whereas, the interviews in the second phase were transcribed before being coded and analyzed. Though coding the interviews in the audio form was simplified by using Atlas ti, it was more convenient to code the written forms of the interviews for easier and better access of the coded extracts. As a result, all the interviews in the second phase were transcribed before being analyzed.

The term "coding" though is widely used; it usually entails different procedures that sometime authors do not explicitly describe (Richards & Morse, 2007). It is essential for the enhancement of the quality and validity of any study to delineate not only the procedures followed in data collection but also in data analysis (Creswell, 2011; Maxwell, 1992; Mishler, 1990). Therefore, the coding steps followed in analyzing the interviews and focus groups transcripts are listed below.

The transcripts and audio interviews were uploaded to the programme in two separate files for coding. Topic coding which links the ideas to the data rather than labeling the data only (Richards & Morse, 2007) was implemented. A list of 20 to 22 codes emerged from reading the transcripts and referring to the study questions. The codes were selected based on their re-occurrence of the ideas and relativity to the study questions. Once the list of codes was refined, the steps below, which were adapted from (Miles & Huberman, 1994), were applied to the interviews and focus group scripts/audio materials (i.e. teacher interviews in Phase 1). The same steps are displayed in Table 4.11 with examples from Phase 2 teacher interviews. These steps were:

a. Assigning codes to the appropriate extracts in all interview scripts or audio recordings.

b. Reading the extracts linked to each code and clustering them into groups.

c. Looking for possible themes.

d. Comparing and contrasting the themes within the same phase and between the phases.

e. Splitting or combining themes.

f. Building a logical chain of evidence.

g. Making conceptual coherence.

Table 4.11. The Process of Coding and Analyzing the Teachers' Interviews in Phase 2

Procedures **Codes**	**Assigning Codes to the Interview Scripts**	**Clustering into Groups**	**Looking for Possible Themes**	**Splitting or Combining Themes**	**Building a Logical Chain of Evidence**	**Making Conceptual Coherence**
Low FP cut-off mark	"Quite honestly some of them do not handle it well but some of them can handle it well. The reason being is that they go from the foundation year to the first year and this is something I have a bit of issue with, having to get only 50% of the total mark."	FP criteria was lenient	Teachers views of the FP in retrospect	Assessment and curriculum in the FP were not appropriate + Gender differences in FP performance (*This belongs to a different theme but was combined here for plausible relatedness with the above theme*)	Gender differences in performance ⇩ Curriculum not suitable for assessment ⇩ Criteria was too lenient	Most First Year teachers tended to express their dissatisfaction with the FP because the curriculum used was not appropriate for FP assessment. Also, several teachers linked FY achievement difficulties to the gender difference in language proficiency. That is, they felt that more male students faced academic challenges because of their lower language proficiency compared to the female students.
FP criteria was lenient	"I think that the foundation selection criteria should be stricter. I used to get a shock when I got some students who cannot even write one sentence right."					

FP criteria was not appropriate	"So foundation have to have and put more emphasis on writing to prepare them for the first year because writing is a big component in year1 assessment and constitutes 30 % of it."	FP curriculum was not appropriate				
Difficulties in English language courses	"I thought that their English would be better given that they were in the first year, so I used different vocabulary and terminologies, then I noticed that they really did not know anything about what I was talking about. Of course only some of them not all of them."	Students faced difficulties in FY courses	Not ready for FY in relation to language and study skills	Examples of students' of inability to handle FY study	Students not ready for FY ⇩ Students struggle in FY courses ⇩ Students lack language and study skills	According to most FY teachers, students in the FY are not ready in relation to both language and study skills for the requirements of the FY courses. These skills include, but are not limited to, poor communication skills, inability to digest lectures, inability to read assigned materials and other poor study skills such as summarizing and note-taking.
Difficulties in academic courses	"Every semester I teach close to 100 students, I can say that 70 to 80% of them have really poor English. They are poor in terms of grammar spelling and in terms of					

	conversing, you know."					
Language skills	"Some of them, I cannot say all of them, but some of them really need more support in English language. Maybe some of them are good in reading but they need to know how to listen to the teacher better and take notes."	Students lacked required skills for the FY		Students lack of language and study skills		
Study skills[12]	"It is the pronunciation of the vocabulary in the presentations. In the essays, they have a problem of grammar, sentence structure and organization of essays. We also have a problem with documentation in the essay; they do not know how to do that. And referencing."					

[12] Some quotes can be categorized under more than one category. The example categorized as "Difficulties in academic success" can also be categorized as "language skills".

4.7.3. Descriptive and Inferential Statistics in Analysing the Questionnaire and Student Scores

The statistical techniques implemented in analysing student and teacher questionnaires and student scores varied based on the purposes of study from using questionnaires and collecting student scores. Therefore the discussion below is divided into two sections, the first is about statistical analysis of the questionnaires and the second is about the statistical analysis of student scores.

4.7.3.1.Statistical Analyses Used with the Questionnaires

The questionnaires were used to accomplish two objectives (1) to provide broader but concise information about student and teacher perceptions of assessment on the FP and in the FY, and (2) to find out if there were perception differences amongst the groups of gender, college, specialization and self-evaluations as some studies suggested (e.g., Cheng, 1999; Cheng, Andrews, & Yu; 2011, Huong, 2001; Xu, 1991). To meet these objectives, the student and teacher responses were analyzed in three stages. Firstly, the frequencies and means of the responses to the five point likert scale were obtained for each item. This means that in the questionnaires, the responses for each item were accumulated in each of the scale's categories: *strongly agree, agree no opinion, disagree and strongly disagree*. This step provided simple and detailed information about how the participants responded to each item in the scale. As was mentioned earlier, the components of each topic were sometimes re-statements of one idea, but they represented different yet related ideas at other times. Therefore, it was vital to look at how the participants responded to each item before separately combining them to represent a whole topic.

Secondly, the Mean and Standard Deviation were calculated for each topic in the questionnaires. The Mean and Standard Deviation values provided concise descriptions of the student and teacher responses to each topic, and facilitated presenting, comparing, discussing and linking their perceptions to concepts and theories in the literature. Thirdly, the student responses were tested for significant differences amongst the groups of gender, college, self-evaluation and specialization, and the teacher responses were tested for statistically significant differences amongst

the groups of gender, college, experience and specializations. Two non-parametric tests of significant differences (i.e. Mann-Whitney U Test & Kruskal-Wallis Test) were used because the responses were skewed and the sample sizes amongst the groups were not equal. These two tests were used to explore the difference in means between groups of the same sample (Dancey & Reidy, 2004). This study was an exploratory study that did not state hypotheses about expected significant differences amongst the groups but sought to understand the implications of any found differences and compared them to previous studied and comparable findings.

4.7.3.2. Statistical Analyses Used with the Student Scores

As pointed out earlier in section 4.3 part of this research project focuses on a correlational study of the predictive validity of English language assessment on the FP (for more discussion on correlational analysis in applied linguistics, see Dornyei, 2007, pp.223-241). It studied the correlation between students' English language proficiency on the FP (measured by their scores in the two English language FP assessment) and their academic achievement in the FY (measured by their average scores in the first semester of the FY assessment). It also focused upon whether the strength of the correlation was influenced by the different groups of students or teachers. Two types of statistical analyses were applied namely correlational analysis using Spearman's rho and the difference in means analysis using Mann-Whitney U test and the Kruskal Wallis Test. The latter tests were used to identify significant differences between student scores in different groups when the predictive validity varied amongst the groups.

The reasons for using non-parametric tests are similar to the ones mentioned in the previous section. The distribution of the scores was negatively skewed and the sizes of the group samples were not equal.

The students' grades were analysed for the most appropriate FP cut-off point (i.e., the score that a student should get to be allowed to take courses in the academic programmes). The students' grades in FP and FY were cross tabulated and a grade of 2.00 was taken as an indication of success in FY study. Extended discussion of the procedure used in the cut-off point analysis is presented in Section 10.34.

4.8. The Quality of the Research Study

Usually, qualitative research designs address issues on reliability and validity differently from quantitative designs; the latter involves terms such as stability/consistency, internal reliability and inter/intra-observer consistency, face, concurrent, predictive, construct and convergent validity (Bryman, 2004). Maxwell (1992) argues that validity in qualitative research is better expressed as "understanding" that is divided into descriptive validity, interpretive validity and theoretical validity stressing that their roles in qualitative studies is marginal compared to their roles in quantitative studies. He also outlined the meaning of generalizability and evaluative validity. The boundaries between these types are blurred and evidence collected for one of them may ultimately provide evidence that supports the unitary nature of validity. Maxwell claims that the lack of reliability is a possible threat to descriptive validity that results from inconsistencies in the data produced.

On the other hand, Mishler (1990) introduces “trustworthiness” to represent validity in qualitative or "inquiry-guided" research. He explains this term as a validation that promotes ongoing appraisal of claims made and a functional criterion of reliability of the findings in further work (p.419). He refers to reliability, falsifiability, and objectivity as methods of supporting validity claims rather than "abstract guarantors of truth". Studies are validated through evidence that "contain within themselves the criteria and procedures for evaluating the 'trustworthiness' of studies and serve as testaments to the internal history of validation within particular domains of inquiry" (p.422). In this study, I intend to make use of both meanings of validity and reliability in qualitative and quantitative methods where suitable in a complementary manner. For example, the data gathering process and the data itself will be provided for scrutiny by other researchers as a way to support the descriptive and interpretive validity. Theoretical validity will be linked to the notion of "trustworthiness" in providing evidence from the literature and respondents' accounts to support the claims or arguments presented. External, internal and construct validity issues will be considered using appropriate methods (see Creswell, 2003, pp. 171-175).

Generally, qualitative research studies do not claim a high generalizability of their findings as the goal is to focus on certain special phenomena in a specific context (Maxwell, 1992). However, the aspect of internal generalizability "within a community, group, or institution studied to persons, events, and settings that were not directly observed or interviewed", is vital (Maxwell, 1992, p.293). Including a quantitative research and triangulating methods of data collection both strengthen and enhance the generalizability of mixed-methods research. Still there are other considerations that should be realized when attempting to generalise in a mixed-method case study research (see Hamersley, 2002).

Reflexivity, on the other hand, is crucial in qualitative studies and it is "the process of recognition of the role of the researcher in co-producing psychological knowledge stands" as Langridge et al. (2007, p.59) stated. Given that the researcher has been involved in the context of the study as a former employee, personal and functional reflexivity will be considered and used to better enhance discussing and presenting the findings in this study.

4.9. Ethical Considerations

According to the principle based approach to research ethics, a research study should be autonomous, non-harmful, beneficial, and just (Wiles, Heath, Crow & Charles, 2001). This study was designed to best meet these requirements in that the participation was voluntary, participants were treated equally in the process of conducting the study, no harm was caused or intended to be caused in the dissemination of the study results, and the results will hopefully inform better decision making. Vaus (2002) added confidentiality, anonymity and privacy to the previous list. These issues were covered by using two consent forms, the first form on gaining permission to conduct the study in CAS was sent to the General Director of CAS in the Ministry of Higher Education. When the first permission was granted, another consent form was handed out to invite the teachers and students to participate. The form gave a brief description of the study and stressed that the data would be kept confidential and anonymous (see appendix 4.2).

The researcher recognizes that the nature of the topic, "assessment practices", led some participants to react apprehensively to the study because of possible fears of being identified or later being questioned on their accounts by authorities. For example, one of the Programme Directors in CAS responded to my e-mail about piloting the questionnaires apprehensively in March 2010 and asked about how the data collected and findings would be used. I explained to him/her the nature of the research and provided him with reassurances on the privacy and anonymity of the participants' identities; eventually he/she helped in piloting the questionnaires. Also, the topic of tests and interpretation of scores is usually surrounded with thorny ethical issues; Dornyei (2007) argues tests as one of the sensitive research aspects that should be considered, he maintains that "we need to note that the misuse of test scores carries real dangers" (p.66). To allay any such fears, the researcher met the participants face-to-face and explained the content and purpose of the study in a manner believed not to risk exposing how the researcher would work or analyze the data. Appendix 4.8 shows a check list adapted from the website of the College of Humanities and Social Sciences of the University of Edinburgh, about the ethical issues that a researcher should consider; which was applied to this research plan in 2010 and which principles have been followed since then.

4.10. Chapter Summary and Conclusion

This chapter has discussed the research design and outlined the structure of this study. The pragmatic view of research prioritizes the questions of a study and considers them as the starting point for selecting and planning research methodology. It was deemed that a mixed-method research was the most suitable type of research considering the study questions. Studies on programme evaluations usually employ a diverse set of methods that investigate not only the effectiveness of the programme but its consequences, as has been clarified in Chapter 3. Following this approach, the current study explores the FP assessment effectiveness and predictive validity through both document and statistical analyses as well as through investigating perceptions using questionnaires, interviews and focus groups. The content of the data collection methods was based on related literature and adapted to the context of the study. Given that this study implemented a number of different methods, various data analysis approaches were used to examine the data generated. These analysis

methods include thematic content analysis and statistical procedures. In mixed-method research, the chances of getting ‘unanticipated’ findings are high.

> Research of all kinds has the capacity to offer suprising or unexpected findings, but when quantative and qualatitive research are combined the possibilities of unplanned or unanticipated outcomes are magnified considerably (Bryman, 2006b, p.124).

Therefore the following seven chapters present the findings based on the data collection methods by which they were generated.

Chapter 5: Document Analysis

5.1. Introduction

This study utilises a mixed-method approach to explore the areas of interest, i.e., multiple types of data were produced and analysed. The organisation of the results chapters is based on the methods used to generate data. This has been done to mark the distinctiveness of the data types, and process of its analysis, and to ease comparison of the findings.

This chapter presents and discusses the results obtained from the document analysis conducted in the first and second phases of the study. It aims at responding to three main questions and four sub-questions as displayed in Box 5.1.

Box 5.1. The Questions Addressed by Document Analysis*

1. How well did the process of assessing students' English language performance, through continuous assessment and tests, function in the Foundation Programme (FP)?
1.1. What processes and procedures were followed in writing and implementing the assessment instruments, as depicted by the official documents?
1.4. What were the differences between the 'continuous assessment' model used in the Academic English Skills course and the 'test' model used in the General English Skills course in terms of effectiveness, accuracy, and preferences of teachers and students?
1.6. What types (criterion/norm-referencing) of assessment were used? And how?
2. How did the assessment instruments correspond to stakeholder wishes?
2.5. What were the national and international policies on teaching and assessing language that influenced assessment in Oman? And how does FP assessment correspond to these policies?
3.4. How demanding were the learning outcomes and assessment of the academic courses in the First Year (FY) of students' language skills?

*Original numbers of questions used as appeared in Chapter 1

5.2. Background on the Role of Documents in the Foundation Programme

The documents analysed in this chapter vary in type, length, accessibility and implementation. Most of them were centrally issued by the Directorate General of the Colleges of Applied Sciences (CAS), some were issued by the Oman Academic

Accreditation Authority (OAAA), and others by the Ministry of Higher Education. The types of documents can be categorised in terms of their focus into general documents, teaching documents and assessment documents. About 118 documents were investigated in this study, varying in length from one page to 50 pages. The following table displays a sample of these documents.

Table 5.1. A Selection of Documents Relating to Teaching and Assessment of the FP English Language Course*

Type	Document Titles
General	Oman Academic Standards for General Foundation Programmes
	Colleges of Applied Sciences: Academic Regulations
	Student Guide for Colleges of Applied Sciences (2011/12)
	Academic Audit Reports on Colleges of Applied Sciences in Sohar, Ibri and Salalah
Teaching	Foundation Programme: 2010-11
	Course Specifications for Foundation English
	Headway Academic Skills (Level 2)
	Headway Plus (Intermediate)
	Essay and Presentation Guidelines
	Foundation Year Academic Calendar
Assessment	CAS English Department Assessment Handbook
	Foundation Year – Level A
	Academic Skills Project & Presentation Topics
	Mid-term and Final Tests for Level A Foundation English
	Assessment Policies: English Department October 2011
	English Department Anti-Plagiarism Procedures: Student plagiarism V3, 02/11
	Marking Scales for Tests and Projects

* The complete list of analysed documents in Appendix 5.1

The accessibility of these documents to FP teachers depends on their position and their target audience. Some of the general documents were accessible to the heads of departments, but not the teachers; others were accessible to all and could be retrieved from the Internet. The general documents could be claimed to be unnecessary for the teachers as they mostly included policies, regulations or audition reports, and consequently, they were not distributed to teachers, though they were available online. The teaching documents were intended to be supplied to every teacher on the FP. It was the responsibility of the course coordinators in each college to supply the teachers with these documents, which were exclusively accessed online by the

coordinators. This means that the number of teaching documents the teachers received was bound to how much and how widely a coordinator disseminated these materials. One coordinator, interviewed in the first phase, expressed a concern that not all of the teaching materials were accessible to the teachers because of the high rate of teacher turnover, despite her persistent efforts to keep all of them well-informed. Similarly, circulation of the assessment documents depended on the assessment coordinators at the colleges who had exclusive online access to these materials. All of the documents on assessment tasks, specifications and marking scales were supposed to be shared with the teachers. Current and previous tests however, were accessed by the assessment coordinators only, to allow a possible recycling of the test tasks as was justified by the programme director.

The level of teacher participation in and implementation of the FP English course documents also differed according the document types. In general, not all teachers participated in writing the documents, including the tests and assessment tasks. Only the assessment coordinators, who taught a lower number of hours, participated in writing the tests. In regard to the implementation of policy documents and marking scales, there was no accountability system in place. However, there were standardisation workshops held for marking the writing task of the General English Skill (GES) final test, and a two-rater policy was followed in evaluating the students' speaking skills in the GES interview; no similar workshops were conducted on the standardisation of marking the Academic English Skills (AES) assessment.

In carrying out the document analysis, I was trying to understand in a factual way the plans and intentions and was deliberately using a problem centred approach to find possible contradictions, some of which might be reflected in the other kinds of instruments.

5.3. Results

The results are categorised into five main themes: (1) conflicts and tensions between criterion-referenced and norm-referenced assessment, (2) compatibility between what was taught and what was assessed, (3) inconsistency in implementing assessment criteria, (4) replication of the academic standards in the FP course

specifications, and (5) language requirements of FY academic courses. These five themes encapsulate an evaluation of the English language assessment on the FP from micro and macro perspectives. The first, second and third themes focused on the design, implementation and marking of the assessment tasks respectively (i.e., a micro perspective). The fourth and fifth themes focused on the evaluation of FP assessment in the context of the national standards of the FP in Oman and its suitability for the language requirements of the FY academic courses (i.e., macro prospective). These themes emerged after implementing the coding process explained in section 4.7.1.

5.3.1. Conflicts and Tensions between Criterion-Referenced and Norm-Referenced Assessment

As explained in the previous chapters, the English language components of the FP consisted of two courses: AES and GES. At the time of this study, GES assessment included a midterm test and a final test that were centrally written, whereas AES assessment included report writing and an oral presentation of the report. Investigation of the official documents on constructing the GES tests appeared to show that there was a sort of incongruity among different official documents about whether the purpose of these tests was norm-referencing or criterion-referencing. For example, the test writing instructions in the *English Department Assessment Handbook* (2010) advised using what could be considered norm-referenced techniques in writing test items and analysing student scores. However, the *CAS Regulations*, *General Foundation Programme Standards (GFPs)* and *English Department Course Specifications* all stated that the tests should aim at assessing students' abilities to achieve set outcomes and, should be using criterion-referenced achievement tests. The policy documents of the Colleges and of the national accreditation institution namely *CAS Academic Regulations* and *Oman Academic Standards for General Foundation Programs*, clearly mandated that assessment instruments should have the traits of a criterion-referenced assessment not a norm-referenced one. This is explicitly stated in the extracts below.

> Normally a final grade in any given course is based on continuous evaluation of the achieved Learning Outcomes. This implies therefore that assessment is determined more by the fulfillment of stated criteria

> rather than by solely comparative achievement within a class (CAS, 2010a, p.15).
>
> All assessment shall be criteria based (i.e., based on the learning outcome standards) and not normative references. Arbitrary scaling of results (for example, ensuring a certain percentage of students passes by moving the pass/fail point down the scale of student results) shall not be permitted (OAAA, 2009, p. 8).

However, the English department's documents seemed to give conflicting guidance. Although, these documents stated that the tests aimed at evaluating students' mastery of a set of learning outcomes, and thus implied that they should be criterion-referenced, the test writing and analysing instructions entailed using norm-referenced methods that compared the students' performances to each other, as in this extract:

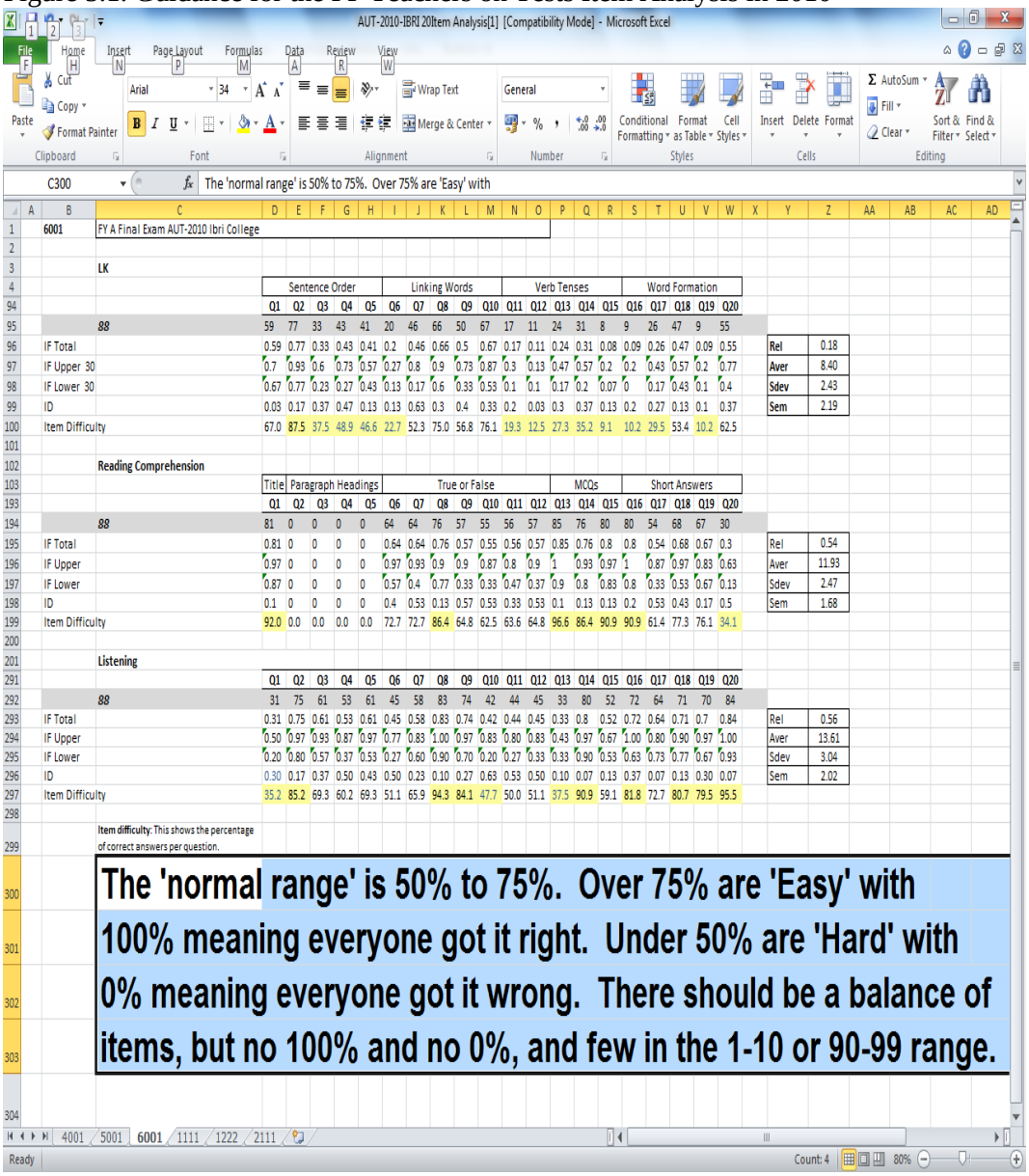
> Item analysis will be carried out by the Assessment Team based on samples of marks from a single college. This analysis involves counting the numbers of correct answers given for each item by the sample population. From this analysis a number of conclusions can be drawn:
>
> - Items which nobody gets right or items which everybody gets right are to be marked for deletion or alteration in subsequent versions of the test.
>
> - Items where 25% or less of the population gets the correct answer need to be investigated: if the 25% of the sample getting the answer right are also the 25% highest scoring students, this is a positive indicator. If no such correlation is found, the item needs to be marked for deletion or alteration in subsequent versions of the test … Such items should be recorded to build up a bank of bad test items in order to guide future test writing (CAS, 2009, p.20).

This was also apparent in the following instructions in the newer version of the same document:

> Preliminary analysis of marks: This should include (a) a check on relative scores for representative students i.e., students who are recognised to be high-achieving, middle-range, low-achieving. If these students are placed in 12 more or less the order teachers would expect, this is a positive indicator (b) a check on relative scores for groups. Again this relates to recognised prior achievement: if groups perceived to be achieving at the same levels score roughly the same, this is a positive indicator (CAS, 2010c, p12).

Also, Figure 5.1 shows that the process of item analysis focuses on selecting the test items using the normal distribution curve, to ensure that most of the population fall in the middle range of the distribution.

Figure 5.1. Guidance for the FP Teachers on Tests Item Analysis in 2010



Though the GES tests did not comply with CAS or OAAA policies on implementing criterion-referenced tests, they did follow the policies on testing achievement, not proficiency. It is stated in the *English Department Assessment Handbook* (2009, p.3) that "the purpose of the test is to show achievement". Hughes (2003) says that achievement tests "establish how successful individual students … have been in

achieving objectives" (p. 13) and identifies the aim of proficiency tests to be "measure[ing] people's ability in a language regardless of any training they may have had in that language" (p.11). It seems that CAS students were generally assessed on a predetermined set of outcomes rather than on general proficiency in certain skills or abilities, as the policy makers intended.

On the other hand, the AES assessment instruments seemed to be designed to evaluate the students' language abilities using criterion-referenced and achievement measures as recommended in CAS regulations and OAAA standards. This was deduced from reviewing the specifications of the AES report and presentation that assessed FP students based on their achievement of a certain set of criteria, and was also expressed in the following extract.

> Continuous assessments are designed to provide teachers and students with an on-going measure of achievement so that they can both adjust expectations and level of input (CAS, 2010c, p. 4).

5.3.2. Compatibility between What Was Taught and What Was Assessed

By comparing and contrasting the focus of assessment instruments with the focus of the taught materials, this section aims not only to facilitate understanding the structure of assessment in the two FP language courses but also to place the students' and teachers' views in a clearer context. It sheds some light on what was claimed to be assessed and what was actually assessed in each course by comparing textbooks, course specifications, test specifications and papers, and continuous assessment specifications and tasks. This part of the study followed an objective based model of evaluation which investigates if the objectives of a programme have been met (see Section 3.3.4).

The following table displays the textbooks and assessment tasks used in each course. It can be seen from the table that GES assessment consisted of tests, while AES assessment consisted of performance assessment tasks[13] (i.e., a report and presentation).

[13] As has been explained in Chapter 2, the term *Test* was used to refer to standardised testing which was the main instrument used in GES assessment. The terms *Performance Assessment* and *Continuous Assessment* are both used to refer to AES assessment which includes tasks such as writing a report or presenting a topic.

Table 5.2. Textbooks and Assessment in AES and GES Courses[a]

FP Course	Textbooks	Assessment Components		% of Course Total	% of English FP Total
GES	*New Headway Plus Intermediate* & *New Headway Plus Intermediate Workbook*	Midterm Test	Language Knowledge	10%	50%
			Reading	20%	
			Listening	20%	
			Speaking	20%	
			Writing	30%	
			Total	**40%**	
		Final Test	Language Knowledge	10%	
			Reading	20%	
			Listening	20%	
			Speaking interview	20%	
			Writing	30%	
			Total	**60%**	
AES	*New Headway Academic Skills (Level 2)*	Presentation		**50%**	50%
		Report		**50%**	

[a] taken from (CAS, 2010b, p.19)

5.3.2.1. Compatibility in GES Learning Outcomes, Taught Materials and Test Tasks

Analyses of GES and AES documents are presented separately. First the GES course materials, textbooks, tests, and scales were examined to understand what the students were supposed to be taught and what was supposed to be included in the tests according to official documents. An initial comparison of the intended GES learning outcomes, as stated in the *Course Specification for Foundation English,* and the GES test specifications, as stated in the *English Department Assessment Handbook*, revealed a very close resemblance, suggesting that most of the skills the students should master by the end of the course seemed to be measured by the tests, if the students' met the specifications (see appendix 5.2 & 5.3). For example, the *Course Specification for Foundation English* stated that "by the end of the course, students should be able to read texts of up to 600 words, with a Flesch test readability score of 85%, with gist, main points and detailed comprehension" (2010c, p.16). This objective was found to be addressed in the *English Department Assessment Handbook*, which stated that the reading passage used in the final test should be "500-550 words of length and of around 80% of readability" (2010c, p.20). From this

example and several others, it can be inferred that the GES test specifications seemed to correspond to the learning outcomes by using tasks of appropriate levels. It can also be suggested that since GES test tasks focused on covering most of the learning outcomes, GES tests fulfilled the requirements of content validity (i.e., the extent to which a test represents all facets of a content domain).

Despite the general compatibility between the course learning outcomes and the test specifications, an analysis of the GES course textbook (i.e., *New Headway Plus Intermediate)* showed that its content, especially its tasks, were of a shorter length than those suggested by the course learning outcomes and test specifications. For example, the reading scripts provided in the textbook seemed to be significantly shorter than the 600 word passages used in the test. Also, the course specifications stated that students should be able to produce 350 word written scripts, yet the writing tasks in the textbook were based on shorter passages. This suggests that the students possibly lacked sufficient and appropriate input to meet the test tasks' requirements. The taught materials were of a shorter length than of that stated in the course learning outcomes and test specifications.

That being said, most of the general topics mentioned in the GES textbook (e.g., talking about films, and cities) were systematically similar to the topics the learning outcomes and test specifications addressed. This was true for each of the reading, writing, and speaking skills, but not for the listening skill.

Although the assessed learning outcomes of the listening skill matched those of the textbook, the test specifications introduced an unfamiliar listening genre to the students (i.e., listening to lectures). The test specifications stated that two listening tasks should be used: (1) a dialogue between two people, and (2) a lecture. However, the lecture genre did not occur in either the textbooks or the listening skill learning outcomes of FP course specifications. Listening to a lecture could be more difficult for the students as a genre; it is a monologue which usually lacks social interaction cues. Though some might argue that this type of listening task is more authentic, it is different to what the students were taught in class (e.g., discussion, role-play and

description) and perhaps more complex. After the midterm test was administered in Spring 2011, the issue of the listening task difficulty came up in several focus groups (see Section 7.2.2.2). Likewise, the difficulty of the listening component of the test was not expressed only by the students, it was also acknowledged in the *English Department Assessment Handbook*, "listening is the most difficult task for students" (2010c, p.8). This reoccurrence of instances where the listening tasks were deemed to be difficult for the students implies a consensus on the inappropriateness of the listening task level or type.

5.3.2.2. Learning Outcomes, Taught Materials and Assessment Tasks in the AES Course

As in the case of the GES tests, the specifications for the report and presentation task used in the AES assessment closely mirrored the intended AES learning outcomes, but again the assigned textbook seemed unable to fulfil the ambitious stated specifications of the assessment and learning outcomes. The learning outcomes in the *Course Specifications for foundation English* included statements such as, "produce a written report of a minimum of 500 words" (2010b, p.19), and "read an extensive text of around 1,000 words broadly relevant to an area of study and respond to questions that require analytical skills, e.g., prediction, deduction, inference" (2010, p.19). However, the course textbook, *New Headway Academic Skills (Level 2),* included reading passages of a maximum length of 600 words and assigned writing activities of 250 word essays. A comparison of the language difficulty levels of the textbook materials and those of the learning outcomes and test specifications reveals considerable differences between them indicating that test specifications might generate test tasks of a more difficult level than those experienced by students in the classroom.

In order to understand the nature of what seems to be assessed using performance based tasks (e.g., a report and a presentation), studying the tasks alone was not enough. The marking scales had to be considered too as they determined the focal points of an assessment through the criteria used. In this study, the band descriptors of the AES learning outcomes and of the marking scales were compared and a

discrepancy was found between what was intended to be taught and what seemed to be assessed. Interestingly, this discrepancy was found only between the *writing* learning outcomes and *writing* marking scale descriptors but not between *speaking* learning outcomes and the *speaking* scale descriptors. Before a fuller description, it seems necessary to first clarify the nature, structure and specifications of the AES assessment tasks: report and presentation. Box 5.2 displays the instructions which teachers were supposed to share with their students on the AES assessment.

Box 5.2. Instructions for Report Writing and Presenting in AES Course[a]

- Students are required to complete a project which involves some library, Internet and real-world research (e.g., interviewing people), a presentation and a report.
- Students should choose a topic from the list below [the list was attached to the instruction sheet]. The topics are based on the subjects the students will study this semester.
- The subjects are quite wide so the student and teacher should agree the actual scope/title of the report.
- Students should not write about Oman or Omani related topics. As part of their project they are required to do research about a *new* topic.
- The report should be around 500 words and the presentation should be at least 5 minutes. Each part represents 50% of the marks.

[a] From (English Department, 2011, p.1)

The focus of the scale used to mark the written report was found to be different from that of the writing learning outcomes of the AES course; these differences were apparent when the learning outcomes of the AES writing skills were placed next to the highest level of the writing marking scale as shown in Table 5.3. It can be seen from the table that four of the six criteria in the scale evaluated the structures and procedures of writing an essay (i.e., word count, plagiarism and implementing suggested changes). All of these four italicised criteria correspond in focus with two learning outcomes of the writing skill in the left hand side of the table. In the scale, there were only two criteria that focused on the content of the report, namely the fifth and sixth points: "addresses chosen topic directly" and "essay structure used includes introduction, conclusion ...etc." Areas such as linguistic knowledge (e.g., using pronouns or modal verbs), and stylistic knowledge (e.g., using paraphrases) were listed in the learning outcomes but were overlooked by the marking scale. It can be

inferred from the marking scale that regardless of the quality of a written piece, a student could easily score a high score if he submitted on time, his report was within the word limit, he wrote it by himself, and he followed a teacher's suggestions.

Table 5.3. Comparison of AES Writing Learning Outcomes and Marking Scale Descriptors[a]

Comparisons	Writing Learning Outcomes	The Highest Level of the Marking Scale
Corresponding Areas	• Produce a written report of a minimum of 500 words showing evidence of research, note taking, review and revision of work, paraphrasing, summarising, use of quotations and use of references. • Cite sources according to the APA system.	• *All outlines and drafts completed and submitted on time.* • *Student has actively tried to implement all changes suggested by teacher.* • *Majority of the essay is in the students own words and credit is given when others' work is used.* • *Meets minimum word limits.*
	• Plan and execute a piece of writing by moving through a series of process stages. • Use mind-maps to brainstorm content for writing. • Use linking words to show logical organisation within and across sentences.	• Addresses chosen topic directly; coverage is fairly comprehensive; little irrelevance. • Essay structure used includes introduction, conclusion … etc.
Conflicting Areas	• Proof-read effectively focusing on a range of surface features. • Complete applications forms. • Reformulate phrases from a sentence. • Paraphrase sentences from a text. • Summarise paragraphs from a text. • Use pronouns to avoid repetition. • Use modal verbs (e.g., may, could) and adverbs of possibility (e.g., possibly). Transfer information from graph to text and text to graph.	No corresponding descriptors

[a]See appendix 5.4 for the complete marking scale for writing

As has been noted earlier, the speaking learning outcomes in the AES course closely resembled the presentation marking scale. Table 5.4 displays the similarities between the speaking learning outcomes and the highest level of the speaking scale

descriptors by placing corresponding learning outcomes and descriptors next to each other.

Table 5.4. Comparison of AES Speaking Learning Outcomes and Scale Descriptors[a]

Comparisons	AES Speaking Learning Outcomes	The Highest Level of the Presentation Marking Scale
Corresponding Areas	• Prepare and deliver a talk of at least five minutes. Use library resources in preparing the talk, speak clearly and confidently, make eye contact and use body language to support the delivery of ideas. Respond confidently to questions. • Address questions from the audience. • Plan and conduct a presentation based on information from written material, interviews, surveys, etc. • Tailor content and language to the level of the audience. • Maintain some eye contact with audience.	• Gets the attention of the audience: highlights objectives of presentation • Postures, gestures and movement enhance presentation. • Complete understanding of topic. Clear evidence of independent study. Able to effectively answer any questions on the topic.
	• Outline and define main concepts. • Follow a presentation format. • Use presentation language (discourse markers etc.).	• Presentation well organised with a logical flow of information
	• Achieve the key aim of informing the audience.	• Topic was covered thoroughly and concisely. No important information missed
	• Observe time restrictions in presentations. • Organise and present information in a logical order at a comprehensible speed.	• Reiterates key points: pulls the entire presentation together effectively. • Uses allotted time fully.
	• Speak in a clearly audible and well-paced voice.	• Few pronunciation errors: delivery is clear.
Conflicting Areas	• Make use of audio/visual aids when giving oral presentations. • Invite constructive feedback and self-evaluate the presentation.	• Few grammatical errors; none of which cause confusion. • A wide range of appropriate vocabulary, correctly used.

[a]See complete marking scale in appendix 5.5

In general, the comparison of AES assessment documents revealed instances of what could be regarded as an imbalance amongst the learning outcomes, textbook materials and marking scales in all of the four skills. The learning outcomes of the writing skill were of a higher difficulty level than the textbook writing activities, and

the focus of the writing marking scale differed from that of the learning outcomes. Similarly, the reading outcomes were of higher difficulty level than the reading activities in the textbook; however, there was not any assessment task on this skill in the AES course. The speaking learning outcomes were not covered by the textbook, but they were almost comprehensively represented in the marking scale, unlike the listening ones which were not covered by the textbook and were not assessed.

The attempt to understand how tests and assessment tasks functioned in the GES and AES courses by exploring the larger picture that encompassed the courses' learning outcomes, textbooks, assessment instruments and marking scales showed that what was stated to be assessed did not always correspond with what was actually assessed.

5.3.3. Inconsistency in Implementing Assessment Criteria

The reliability and consistency of assessment instruments in measuring intended English language skills are crucial to effectiveness and validity of language programme assessment. Therefore educational institutions usually record how reliable their assessment instruments are and how consistency in using certain measures should be realised. Accreditation and quality assurance agencies usually urge academic institutions to (1) use reliable measures of achievement, and (2) state the process used to insure consistency in applying these measures. The General Foundation Programmes standards (GFP) as set by the OAAA emphasise the necessity of putting in place appropriate procedures to ensure the required level of moderation and standardisation in language assessment. The extract below addresses Higher Education Institutions (HEIs):

> HEIs must have appropriate internal quality controls for its assessment processes. These must include, at least, internal moderation by faculty of examination papers and of marked work prior to the issuance of results, and a transparent appeals process for students (OAAA, 2009, p.8).

In line with the OAAA standards for moderation and standardisation, CAS regulations included an article on forming a committee responsible for ensuring that standardisation policies within and across the six Colleges are met.

> The aim of the [Examiners] Committee is to:
>
> - ensure consistent standards of quality within the program and across all Colleges, by reviewing the performance for each student enrolled into the program;
> - ensure that all evaluation and grading is performed in a fair and equitable manner, and in accordance with these Regulations (*CAS*, 2010a, p. 15).
> -

The English language Department at CAS, following the guidelines of OAAA and CAS on standardisation and moderation of assessment, issued three policy documents in 2009, 2010 and 2011 respectively. Each of the documents implied that the previous one had fallen short of fulfilling standardisation requirements; it was stated that "unfortunately, this [standardisation] approach has presented severe reliability problems because of varied levels of challenge and it has also meant an excessive workload for coordinators" (CAS, 2010c, p. 5). The changes in the standardisation and moderation policies have been tracked from 2009 to 2012; these changes are listed in Table 5.4 to reflect how the perception of assessment reliability has evolved and how the documents stated it should be realised. The main changes could be summarised in the following six points.

2. In the 2009 and 2010 documents, only the GES assessment instruments (i.e., speaking and writing sections of the final test) were addressed in the standardisation policies. However, the standardisation policies released in 2011 addressed also the AES assessment instruments (i.e., report and presentation) (see row 2 of Appendix 5.6).

3. In the 2009 and 2010 documents, the policies included instructions about two processes (i.e., standardisation and moderation). In the 2011 document, the policies addressed three processes (i.e., standardisation, marking and moderation).

4. The meaning of the concept "moderation" seems to have changed across the 2009, 2010 and 2011 documents to be more about reconciling discrepancies in teachers' scores rather than analysing test items and scores across colleges (see rows

6 and 7 of the same appendix). In the 2009 and 2010 documents, post-moderation was stated to "be carried out by the Programme Director with regard to comparisons of scores between colleges and by the Assessment Team with regard to item analysis". However, in the 2011 document, post moderation was introduced as "discrepancies arising from individual biases are likely to be resolved through reference to a third party".

5. Both the 2009 and 2010 documents acknowledged the English language departments' failure to meet the set principles of standardisation and moderation (see row 1 of appendix 5.6). The 2011 document expected challenges in applying its policies (see row 4 of Appendix 5.6). The failure to moderate marking the AES assessment was also reported in Phase 1 teacher interviews (see Section 7.3.4.1).

6. The 2009 and 2010 documents recommended standardising FP assessment by carrying out workshops where samples of written scripts and oral interviews were marked so teachers would have a feel of what the scores represented before marking the rest of the reports and interviews. The documents, however, did not specify the method of obtaining early samples of the reports and interviews. This point was raised in the 2011 document where the policies advise conducting several presentations and collecting several scripts for standardisation and moderation purposes before commencing with marking all scripts and presentations (see row 2 of Appendix 5.6 for the 2009 and 2010 documents and rows 3 and 4 for the 2011 document).

7.

8. Finally, the 2009 and 2010 documents dealt with the cross college standardisation as a comparison of students' scores in Language Knowledge quizzes and written assessment scripts amongst colleges. The 2011 document addressed the same issue more comprehensively where samples from presentations, reports and speaking tests were required too (see row 2 of Appendix 5.6 for the 2009 and 2010 documents and rows 3 and 4 for the 2011 document).

Regardless of the discussed process of adapting and refining a set of policies for moderating and standardising the FP assessment in and across the colleges, in practice, standardisation across colleges has been limited to the writing section of the GES tests only as has been affirmed by a member of the directing team (personal communication, April 1, 2012). CAS is still struggling to standardise marking the AES assessment tasks.

5.3.4. Replication of National Academic Standards in FP Specification

As the FP is expected to be audited in the near future, its documents (i.e., course specifications, FP handbook, assessment handbook … etc.) intentionally and systematically were designed to adhere GFP standards to the letter. The intention to fully comply with these standards was stated in the *Foundation Programme 2010-2011* document.

> The programme must meet the Oman Accreditation Council's General Foundation Programme Standards. These standards apply to all higher education institutions in Oman, private and public and compliance with the standards is ***mandatory*** by academic year 2010-11(CAS, 2010d, p.1).

The GFP standards provided a set of learning outcomes that could guide HEIs to understand what was expected of a foundation programme. A comparison of these standards with the FP learning outcomes indicated that the standards seemed to be closely followed by FP course specifications, but there were real doubts about how closely (see Table 5.5.). The similarities and sometimes equivalence of FP and GFP's learning outcomes raises doubts about whether the process of writing the Foundation Programme learning outcomes involved any planning or consideration of the unique situation of the students at CAS.

These doubts were strengthened by the fact that the listening and speaking learning outcomes of the AES course were listed in the course specifications with a note saying that they were not covered by the textbooks and teachers should provide appropriate materials to meet them (see Appendix 5.3). Also, in the AES course, the students were not evaluated on the listening and reading skills which are part of the course specifications. This seems to suggest that the writing, reading, speaking and

listening learning outcomes of the AES course were copied from the GFP standards as part of a blind matching process, possibly in order to perform well in the upcoming audition mentioned above.

The underlined phrases in the *CAS English Foundation Course Specifications* (2010) in Table 5.5 are identical to those in the *Oman Academic Standards for the General Foundation Programs* (2008), shown on the right hand side of the table.

Table 5.5. Similarities between AES Learning Outcomes and the GFP Standards

Document / Skill	**CAS English Foundation Course Specifications (2010)**	**Oman Academic Standards for the General Foundation Programs (2008)**
Reading	Read an extensive text of around 1000 words broadly relevant to an area of study and respond to questions that require analytical skills, e.g., prediction, deduction, inference (2010, p.18)	Read an extensive text broadly relevant to the student's area of study (minimum three pages) and respond to questions that require analytical skills, e.g., prediction, deduction, inference. (p.10)
Writing	Produce a written report of a minimum of 500 words showing evidence of research, note taking, review and revision of work, paraphrasing, summarising, use of quotations and use of references (p.19)	Produce a written report of a minimum of 500 words showing evidence of research, note taking, review and revision of work, paraphrasing, summarising, use of quotations and use of references (p.10)
Listening	Take notes on longer talks/mini-lectures (10-15 minutes) (p. 19).	Take notes and respond to questions about the topic, main ideas, details and opinions or arguments from an extended listening text (e.g., lecture, news broadcast). (p. 10)
Speaking	Prepare and deliver a talk of at least 5 minutes. Use library resources in preparing the talk, speak clearly and confidently, make eye contact and use body language to support the delivery of ideas. Respond confidently to questions.(p.19)	Prepare and deliver a talk of at least 5 minutes. Use library resources in preparing the talk, speak clearly and confidently, make eye contact and use body language to support the delivery of ideas. Respond confidently to questions. (p.10)

It can be clearly seen from the table that the AES learning outcomes do not only address similar areas to those of the GFP standards, but are very comparable and identical in language. This finding might explain the mismatch between the focus of

AES textbooks and that of AES learning outcomes, as has been mentioned previously.

5.4. Language Requirements of the Academic Courses in the First Year

The previous three sections addressed issues on document analysis of the FP assessment, while this section attempts to make a link between the language skills focused upon by the FP assessment and the ones required by the FY academic courses assessment. This section seeks a qualitative understanding of the predictive validity of the FP assessment that will be discussed in Chapter 10, by identifying the language skills required in the academic courses. Analysing the relevant course documents (i.e., syllabus, test papers, and course specifications) may help to clarify, add meaning to, or provide a counterweight to the numerical findings on FP assessment predictive validity when considering the specialisations of the students as reported in Chapter 10. Therefore, this section first explores the language demands implied in the FY academic course specifications, and then investigates the linguistic nature of the written output required in the coursework and final tests of these courses.

5.4.1. Comparison of the FP English Syllabus and the FY Academic Courses Syllabi

In order to understand the language focus of the academic courses, the syllabi of the introductory courses of the Information Technology (IT), International Business Administration (IBA) and Communication Studies (CS) were analysed to identify the learning outcomes that seemed to demand linguistic skills. These learning outcomes were compared with those of the FP. Table 5.6 displays the learning outcomes of these courses. Those that seem to require complex English language output in the academic courses syllabi and the FP outcomes that seem to match the academic courses' linguistic demands, are highlighted. One learning outcome is highlighted in the IT course, three in the IBA course, three in the CS course and three in the FP course.

The initial comparison of the highlighted learning outcomes in the FP and FY courses suggests that most of the language skills drawn upon by the academic course outcomes were covered by the FP outcomes. For example, in the FP, students were

expected to master discussing issues in written and oral forms (see points 3 and 6, row 1 in Table 5.6). These two learning outcomes seem to correspond with the linguistic demand of discussing or explaining concepts entailed in the learning outcomes list of the IT and IBA courses. Similarly, the FP learning outcome of being able to read around 1,000 words (see point 2, row 1) could presumably equip the students with the skills needed to understand or identify certain concepts from reading passages as required by all academic courses' learning outcomes.

Table 5.6. The Learning Outcomes of the FP English, IT, IBA and CS Courses

Course	Objectives
FP English (*Foundation English Course Specification*, 2010, p. 18 & p.19).	• Read an extensive text of around 1,000 words broadly relevant to an area of study and respond to questions that require analytical skills, e.g. prediction, deduction, inference. • Produce a written report of a minimum of 500 words showing evidence of research, note taking, review and revision of work, paraphrasing, summarising, use of quotations and use of references. • Take notes on peer presentations, sufficient to enable the student to re-construct the main points of the presentation. • Take notes on longer talks/mini-lectures (10-15 minutes). • Prepare and deliver a talk of at least 5 minutes. Use library resources in preparing the talk, speak clearly and confidently, make eye contact and use body language to support the delivery of ideas. Respond confidently to questions.
IT (*Fundamentals of Information Technology*, 2008, p.1)	• An introductory understanding of computer systems, their components, and their interactions. • Competence with application software, in particular word processing, spread sheets and graphics programs. • An understanding of both why good ergonomic practices are important, and how to apply them in a personal context. • An introductory understanding of the development of the Internet, the World Wide Web, and multimedia; their interactions and common uses/applications, in particular e-commerce. • The ability to discuss the impact of computer technology on society. • An understanding of study paths and career opportunities in information technology. • A broad understanding of ethical concepts related to computing.
IBA (*Bachelor of International Business Administrations*,	• Identify the factors that influence the contemporary business environment. • Discuss the challenges of business, with a focus on the

2008, p.23)	Omani context. • Recognise issues and concerns (e.g., accounting, marketing, finance) related to a current business scenario. • Explain the relationship of business to socio-economic conditions. • Demonstrate an interest to manage an entrepreneurial undertaking.
Communications Studies (*An Introduction to Personal Communication: Student handbook*, 2008, p.2)	• Demonstrate an understanding of the basic concepts involved in the communication process. • Identify the reasons for communication breakdown. • Demonstrate a basic understanding of non-verbal communication cues. • Demonstrate the skills necessary to give a competent oral presentation. • Identify and practise the basic factors involved in effective group work. • Demonstrate an understanding of the cultural factors which have an effect on communications.

However, a comparison of the courses' learning outcomes provides brief information about what language skills were required in the FY academic courses and what was offered by the FP English language courses. These outcomes did not specify the form of the language/non language mediated outcomes that the students were expected to produce; in other words, how these learning outcomes should be realised. Therefore, the assessment used in the IT, IBA and CS introductory courses in the first semester of FY are reviewed in the next section to obtain a deeper understanding of how and what English language skills were required to undertake the FY academic study.

5.4.2. Investigating Assessment of the Academic Courses

5.4.2.1. The Course Work

The types of assessment tasks given to the students in the first semester of their FY study were analysed in terms of the apparent language requirements. Assessment in the IT course was divided into two parts: course work which, as stated in the course specifications, evaluated the practical skills imparted during the course, and a final test which evaluated the students' understanding of the main theoretical concepts introduced in the course. In the IT course, a graphics assignment, lab work and a lab exam were used to evaluate certain IT skills; they used basic language in the instructions and required very limited language responses (see Table 5.7). Similarly, the IBA coursework included a series of multiple-choice quizzes and individual e-learning activities which seemed to require moderate language use, and a final test to

evaluate students' understanding of the focal theoretical concepts and issues. The CS coursework, on the other hand, used assessment tasks that seemed to demand good mastery of the English language, such as an informative talk, a presentation and a 1,000 word essay. It also included a final test to evaluate the students' grasp of the main concepts introduced in the course. The number and weightings of the instruments used in the coursework part of assessment in the academic courses are displayed in the table below.

Table 5.7. Assessment Instruments in FY Academic Courses

Course	Assessment	Weightings
IT	Graphics assignment	15
	Completion of lab work	15
	Lab exam	35
	Final test	35
	Total	**100**
IBA	Quiz 1	20
	Quiz 2	20
	Group assignment, e-learning activities	20
	Final test	40
	Total	**100**
Communication Studies	3-Minute informative talk	15
	1,000 word essay	25
	5-Minute persuasive presentation	20
	Final test	40
	Total	**100**

5.4.2.2. The Final Test

Assessment in all academic courses included a final test, so it seemed feasible to identify some of the language requirements of the output stimulated by these test tasks. The Programme Directors of IT, IBA and CS agreed to provide Spring 2009 final tests for the purposes of this study. The test tasks for each final test were studied in terms of the linguistic nature of the responses they entailed, and samples of these tasks are presented in the table below, organised according to their type and weightings.

The IBA and CS final tests constituted 40% of the total course mark, and the IT final tests constituted 35% of the final mark, but the test papers themselves were designed to be marked out of 50, 100, or 80 respectively; this was then converted to the mentioned percentage of the total course weightings. The IT and IBA test tasks both

utilised the multiple choice and true/false format in the first section of the tests. In the second section, they both used a short questions format that required defining concepts or mentioning elements of a concept, and could be answered by memory and did not seem to involve much original language use (see Table 5.8). The long answer questions used in the third section of both tests seemed not very different from the short ones in terms of the language output they required, as they also focused on reproduction of definitions, discussion of constituting elements in a concept, or explaining reasons for a certain phenomenon. They did not seem to demand any kind of originality of expression, reasoning or thoughtful arguments.

Although the CS test included similar multiple choice and short answer test items, it differed from the IT and IBA tests in using inference and long answer test tasks that seemed to demand additional language skills. The test tasks on making inferences required students to paraphrase and apply previously learned concepts to new contexts; this arguably might need a good command of language to be accomplished. Likewise, the CS long answer test task entailed writing 800 to1,000 words of novel language that should be based on both known or memorised information and individual judgements and thoughts (see Table 5.8). This task required building an argument about the definition of "Intercultural Communication", linking this to a theoretical background and supporting it with examples. Clearly this type of test task required a very good command of the English language.

Table 5.8. IT, IBA and CS Test Tasks Types and Examples from Spring 2009 Final Tests

Course	Test Task Type	Example	Weighting
IT	Multiple choice Question	_____is one of the Arguments for Telecommuting.	30
	Short answer Question	List four benefits of E-commerce to society.	35
	Long answer Question	Briefly explain what is meant by "a system" and give three examples of systems.	15
Total			**80**
IBA	True/False Question	In the Hygiene (Two-Factor) Theory, workers work hard because they expect rewards for a good performance.	10
	Multiple choice Question	Franchised business in Oman is growing. Which is a *franchised* company?	10
	Short answer	Define *culture* and discuss *three (3)* reasons why	10

	question	understanding it is important in business.	
	Long answer Question	*List* and *explain* the five (5) human needs according to Maslow's Hierarchy of Needs Theory.	20
Total			**50**
CS	Inferring from a Reading test	Locate your example by indicating the line numbers and then paraphrase what is being communicated in these lines.	45
	Short answer question	What is conflict? Provide a definition of conflict. Please refer to communication terms (theory) in your answer.	30
	Long answer Question	Intercultural communication refers to communication between people who have different cultural beliefs, values or ways of behaving. Discuss this statement with reference to intercultural communication theory and give specific examples to illustrate these concepts. (800-1,000 words essay)	25
Total			**100**

From the previous analysis of the academic course specifications, assessment schemes and test papers of the IT, IBA and CS introductory courses, it seems that the CS specialisation required a good command of the English language to successfully complete its assessment tasks much more so than the IT and IBA specialisations. Though from the CS course specifications alone this conclusion cannot be decisively made, the types of assessment instruments and test tasks used in the CS course revealed considerable demands on students' language skills.

5.5. Discussion

The four main issues raised above will now be discussed and linked to previous studies. These issues were: (1) conflicts and tensions in using norm and criterion-referencing principles, (2) compatibility between what is assessed and what is taught in AES and GES, (3) inconsistency in implementing assessment criteria, and (4) replication of the GFP standards in the FP specifications. These four areas could be considered as evidence on the content validity of FP assessment.

5.5.1. Norm vs. Criterion-Referencing Tests

Document analysis revealed that the stated intention of using criterion-referenced assessment in the FP was blurred by the actuality of using norm-referencing procedures in GES tests construction and analysis. Policy documents issued by

OAAA and CAS clearly stated that assessment in the FP should be criterion-referenced not norm-referenced. Likewise, policy documents on the FP implied that criterion-referenced assessment was used, yet the GES test writing and analysing instructions in the same documents involved comparing students against each other, which is a characteristic of norm-referenced tests. Bachman (2004) says that aiming at most scores to be around the 50% mark of the test scores range is a characteristic of norm-referenced tests, in which the distribution of the scores should be normal, whilst criterion-referenced tests tend to be negatively skewed showing that most of the students have mastered the course objectives. In this study, it was found that the GES test writing and analysis procedures showed norm-referencing attributes implied in the stated test-writer instructions to compare the students against the low, medium and high groups of achievement. Also, the instructions dictated that the test items with difficulty indices of 0.25 or lower should be investigated for a positive correlation with the high achievers' scores. These procedures are clearly characteristics of norm-referenced tests (Bachman, 2004).

When a test is norm-referenced, mastering the learning outcomes does not become a priority. Consequently, some students can pass the FP without mastering all its stated learning outcomes. Thus, criterion-referenced assessment has been widely enforced by policy makers (Brindley, 2001; Lorena, 2007; Llosa, 2007). Sizmur and Sainsbury (1997, p.129) refer the appeal of criterion-referenced assessment to the need to ensure the "minimal standards in basic skill areas, and the need to produce reliable measurement of these". In line with this view, the purpose of disseminating the GFPs document was stated to "seek to help ensure that those programs (GFPs) are effective in helping students attain the prescribed students learning outcomes" (2007, p.4). Moreover, Sizmur and Sainsbury (1997) argue that criterion-referencing cannot be considered as a trait of a test; it is a concept that is defined by the interpretations made about the test scores and how they are used. If the test was designed to compare students performances against each other and the scores were analysed following the same purpose, then the used test makes norm-referenced interpretations of students English language abilities, thus the test shows attributes of norm-referencing. Applying this understanding to the context of this study, we can

conclude that the GES test interpretations did not conform to the GFP standards when it made norm-referenced interpretations, however, the AES assessment tasks (i.e., report and presentation) made criterion-referenced interpretations.

In discussing the wash-back of English language tests, Shohamy (2007, p.126) points out that "language policy *documents* often become no more than declarations of intent that can easily be manipulated and stand in stark contradiction as the 'tested language' obtains prestige and recognition". A similar argument can be made about the use of norm-referenced tests when actually criterion-referenced tests were recommended by policy documents which were used as "declarations of intent".

5.5.2. Incompatibility between What is Assessed and What is Taught

Several writers in the field of language testing argue that there should be a clear link between what is tested and taught in achievement tests (e.g., Bachman 1990, Fulcher & Davidson, 2007; Weir, 2005). Comparing the documents on assessment specifications, learning outcomes and content of textbooks revealed a clear incompatibility between what is taught and what is assessed. In both AES and GES courses, there were examples of how the intended course outcomes were matched by parallel test tasks, but underrepresented by the course materials. In the GES course both the writing and reading test tasks were at higher levels than the textbook tasks. In the AES course, the incompatibilities appeared in the writing scale used to mark the essays. The focus of the marking descriptors was substantially different from the writing learning outcomes. The descriptors highlighted the procedures of writing and submitting the essay more than the content and language accuracy of the essay. In the AES assessment, the incompatibilities also appeared in the speaking and listening learning outcomes mentioned in the course specification which were not covered by the textbook or assessment tasks. Though Hughes (2003) proposes that achievement tests should be built on stated objectives, not actual teaching, to generate positive wash-back effect, others (e.g., Weir, 2005) argue against this proposition and stress that achievement tests should be based on prior learning experiences not on intended ones. In the present context at least, Weir's view is more pervasive

The above instances of incompatibility suggest a serious issue with the validity of FP assessment. Messick (1996) argues that there are two major threats to assessment validity which he entitles: construct underrepresentation, and construct-irrelevant difficulty. The criteria used in the AES essay marking scale, as shown by the results, underrepresented language accuracy and overemphasised procedures and technicalities of writing such as incorporating teacher comments or submitting on time. Incorporating teacher comments could be a very useful step in the process of writing but it should not be overstressed at the expense of other important language related criteria such as paraphrasing or using appropriate modal verbs and pronouns. Likewise, the GES test embodied features of construct-irrelevant difficulty in the listening task by testing students on an unfamiliar genre. Though some aspects of the AES tasks and GES test showed features of lower validity, it cannot be claimed that they were invalid assessment instruments. Messick advised that compelling evidence from multiple sources should be accumulated to evaluate assessment validity. A more comprehensive discussion of FP assessment validity is presented in Chapter 11.

5.5.3. Inconsistency in Implementing Assessment Criteria

Though the policies of assessment standardisation and moderation were inaugurated in 2009, and were amended in 2010 and 2011, the process of implementing these policies still faced challenges in practice. The main two challenges were identified to be:

- how scripts or recordings for the writing and speaking tasks could be obtained prior to the presentation or essay submission date for standardisation purposes in colleges;
- how cross-college standardisation in marking the writing and speaking component of the assessment could be accomplished.

The *Assessment Polices* document (CAS, 2011) proposed that some of the presentations/speaking tests should be conducted in advance to be used as samples for marking the rest of the presentations. Also, it was suggested that a standardisation session should be conducted after the essays were submitted using a sample of the submitted scripts. All of these measures were intended to ensure consistency in marking the speaking and the writing components of assessment in the colleges, but

they did not address cross-college standardisation. Also, the policies seemed to be suggestions more than commands. The results from analysing the policy document suggest that the moderation and standardisation policies were not all applied in practice.

Similar issues have been highlighted in the literature: Brindley (1998), in a review of studies on outcome based assessment, found that this type of assessment raised concerns about the validity of the descriptors and the objectivity of teachers' judgements. He asserted that empirical studies showed instances of subjective and interpretation-based marking even when the scales were deemed to be clear by the teachers.

5.5.4.Replication of GFP Standards in FP Specifications

Language assessment in education has been affected by the international trend through ensuring accountability in reporting achievement through using outcomes based assessment, as indicated earlier (Brindley, 2001; Llosa, 2007). Llosa (2011) explained that the rationale for standard based reforms was "to improve the quality of education for all students by developing rigorous standards and aligning instruction, assessment, professional development, and resources to those standards" (p.367). Similarly, the FP in Oman is obliged to comply with GFP standards produced by the OAAA. The results of document analysis showed that the FP did not only (on paper) comply with the national standards; its AES learning outcomes actually replicated the ones in the GFP document. The GFP standards were used as the basis for the AES marking scales, not as guiding standards for what should be taught in classrooms. This finding can partly explain some of the students' and teachers' concerns about the difficulty levels of AES assessment.

5.6. Summary and Concluding Remarks

In this Chapter, the findings of thematic analysis of various types of documents were presented under four main headings. The first was how norm-referenced tests were used instead of the criterion-referenced tests mandated by the national and CAS policies on language assessment; it was argued that norm-referenced tests should not

be used in FP assessment as they can have serious negative consequences. The chapter then explored inconsistencies amongst learning outcomes, materials taught, and assessment specifications; these inconsistencies were linked to the blind replication of the GFP standards. The third part revealed difficulties in standardising and moderating marking processes and highlighted inconsistencies in using marking scales, which will recur in the findings from other sources in the following chapters. The fourth investigated the language skills required in FY courses by analysing course specifications, required learning outcomes and actual test papers. This analysis concluded that the CS learning outcomes and assessment instruments, including the final test, seemed to rely on students' language skills more than did the learning outcomes and assessment instruments of the other specialisations.

Chapter 6: The Results of Student and Teacher Questionnaires in Phase 1

6.1. Introduction

This chapter presents the results obtained from the student and teacher questionnaires that were conducted in the first phase of the study in an attempt to answer the study questions listed below (see Box 6.1). It first offers an analysis of the students' responses, followed by an analysis of the teachers' responses, both including (1) the demographic characteristics of the participants, (2) the average responses to the individual items of the questionnaires, (3) an analysis which involves grouping items and looking at the statistical patterns.

Box 6.1. Study Questions Addressed by the Student and Teacher Questionnaires in Phase 1*

1. How well did the process of assessing students' English language performance, through continuous assessment and tests, function in the Foundation Programme (FP)?

1.2. How was the reliability and validity of FP assessment viewed by students and teachers?

1.3. How was the impact of FP assessment perceived by students and teachers?

1.4. What were the differences between the 'continuous assessment' model used in the Academic English Skills course and the 'test' model used in the General English Skills course in terms of effectiveness, accuracy, and preferences of teachers and students?

1.5. How did the teachers perceive the centrally controlled assessment used in CAS?

1.7. In all the above, were there any significant differences between the views of the students' grouping by college, gender, age, self-evaluation and teachers' grouping by gender, college, age, nationality, teaching and assessment experiences?

*The questions numbers are as appeared in Chapter 1.

6.2.The Student Questionnaire

6.2.1. Demographic Characteristics of the Participants

In the first phase of the main study, about 220 FP students were invited to participate in this study, which was conducted over two academic semesters. Of those a total of 184 students participated in responding to the questionnaire; 127 (69%) of them were from Rustaq College and the other 57 (31%) students were from Sur College. The sample consisted of 119 female students (64.7%) and 65 male students (35.3%). At CAS, when the students were admitted to the FP, they had already selected their

intended specializations. The participants were from four different departments: Information Technology (IT), Communications Studies (CS), International Business Administration (IBA) and English Language-Education (see Table 6.1).

Table 6.1. The Distribution of Participants by Specializations

Specialization	***n***	**%**
Information Technology	50	27.2
Communication Studies	25	13.6
International Business Administration	85	46.72
English Language-Education	24	13.0
Total (*N*)	184	100

6.2.2. Students' Responses to the Questionnaire

Table 6.2 displays the number and percentage of the students who responded to each item by selecting a point in the five-point likert scale questionnaire. The points 1 to 5 respectively denote Strongly Agree (SA), Agree, (A), No Opinion (NO), Disagree (D), and Strongly Disagree (SD). The table also shows the means of the students' responses to the items including the recoded ones which are explained below. The mean for each item is calculated by adding up a selected likert point multiplied by the number of participants who selected it then divided by the total number of respondents to that item.

$$\text{Item M} = [(n) \times 1 + (n) \times 2 + (n) \times 3 + (n) \times 4 + (n) \times 5] / N$$

(M) = mean of responses to an item.
(n)= number of respondents to each category in likert scale in an item.
(N)= total number of respondents to an item.

An item mean $1.0 \leq M \leq 2.99$ signifies either that most of respondents agree with a this specific item or that the strength of agreement is more than the strength of disagreement. It is possible that a mean indicates disagreement even when the number of students who agree with a specific item is more than the number of those who disagree, if the number of those who selected 'strongly disagree' is more than those who selected 'strongly agree'.

It is important to note here that the views of the students and teachers on assessment validity are actually a form of "face validity", they are not evaluations of the "construct" or "content" aspects of assessment validity.

Table 6.2. Frequency, Percentages and Mean of Responses to the Student Questionnaire in Phase 1

Topic	**Subtopic**	**Item**	**SA 1**	**A 2**	**NO 3**	**D 4**	**SD 5**	**Mean**	**Mean** (Recoded[14] Items)
1. Perceived Validity	Content	1.1. There is a strong connection between what we do and learn in classroom and the final test.	21 (11.4%)	83 (45.1%)	14 (7.6%)	45 (24.5%)	19 (10.3%)	2.77	-
		1.2. There is a strong connection between what we do and learn in classroom and the continuous assessment.	34 (18.5%)	89 (48.4%)	14 (7.6%)	29 (15.8%)	14 (7.6%)	2.44	-
		1.3. I understand how my language performance will be assessed in FP.	20 (10.9%)	66 (36.1%)	22 (12%)	48 (26.2%)	27 (14.8%)	2.98	-
		1.4. The assessment instruments provide me with enough feedback on my English language performance.	42 (22.8%)	84 (45.7%)	15 (8.2%)	37 (20.1%)	5 (2.7%)	2.34	-
	Construct General	1.5. My scores in language assessment reflect my real achievement level in FP.	56 (30.4%)	106 (57.6%)	9 (4.9%)	7 (3.8%)	2 (1.1%)	1.85	-
	Test	1.6. Tests in FP assist me to function in English language in real life.	43 (23.4%)	75 (41.7%)	12 (6.7%)	39 (21.7%)	11 (6.1%)	2.44	-
		1.7. Tests in FP assist me to function in English language in academic studies.	58 (31.5%)	86 (46.7%)	14 (7.6%)	20 (10.9%)	4 (2.2%)	2.09	-
	CA	1.8. Continuous assessment in FP assists me to function in English language in real life.	47 (25.5%)	103 (56%)	6 (3.3%)	18 (9.8%)	7 (3.8%)	2.04	-
		1.9. Continuous assessment in FP assists me to function in English language in academic studies.	68 (37%)	86 (47.8%)	9 (4.9%)	16 (8.7%)	2 (1.1%)	1.89	-

[14] Recoded means that the responses were reversed to conform to the general meaning of the encompassing topic (i.e., if a response was 1, it was changed to 5 and vice versa).

2. Perceived Reliability		2.1. The scores awarded to the different FP assessment instruments such as: the tests, quizzes, reports and presentations are appropriate.	13 (7.1%)	102 (55.4%)	18 (9.8%)	41 (22.3%)	4 (2.2%)	2.56	-
		2.2. Usually the difference between my scores in the tests and continuous assessment is not considerable.	14 (7.7%)	57 (31.1%)	23 (12.6%)	75 (41%)	14 (7.7%)	3.1	-
3. Preference of Tests		3.1. I would prefer to have a final test only instead of continuous assessment and a final test.	39 (21.2%)	79 (42.9%)	23 (12.5%)	26 (14.2%)	16 (8.7%)	2.46	-
		3.2. Some sections of the continuous assessment should be changed.	52 (23.4%)	74 (47.3%)	24 (12.0%)	23 (13.6%)	8 (3.3%)	2.26	-
4. Preference of CA		4.1. Some sections of the final test should be changed.	96 (52.2%)	69 (37.5%)	5 (2.7%)	8 (4.3%)	5 (2.7%)	1.67	-
		4.2. The continuous assessment provides me with a better opportunity to show my English language skills compared to the tests.	43 (28.3%)	87 (40.2%)	22 (13.3%)	25 (12.7%)	6 (4.4%)	2.23	-
5. Satisfaction with current assessment tools		5.1. I am satisfied about the types of assessment instruments used to evaluate my English language skills.	21 (11.4%)	128 (69.6%)	18 (9.8%)	12 (6.5%)	2 (1.1%)	2.15	-
		5.2. FP assessment instruments should not have fewer different parts (tests, presentation, written report, quizzes ...etc.).	22 (12%)	30 (16.3%)	14 (10.3%)	70 (38%)	40 (21.7%)	3.42	-
		5.3. The assessment instruments should be changed to include aspects of students' English language that are not assessed currently. (recoded)	58 (31.5%)	86 (46.7%)	10 (5.4%)	17 (9.2%)	8 (4.3%)	2.01	3.85

6. Perceived Impact	social	6.1. Tests in the FP make me feel stressed. (recoded)	58 (31.9%)	66 (35.9%)	15 (8.2%)	33 (17.9%)	10 (5.4%)	2.29	3.70
		6.2. CA in FP makes me feel stressed. (recoded)	32 (17.4%)	65 (35.3%)	13 (7.15)	38 (20.7%)	7 (3.8%)	2.1	3.43
		6.3. English language assessment in FP is fair.	31 (16.8%)	72 (39.1%)	18 (9.8%)	43 (23.4%)	20 (10.9%)	2.72	-
		6.4. English language in FP assessment is not frightening to me.	83 (45.1%)	53 (28.8%)	13 (7.1%)	18 (9.8%)	14 (7.6%)	2.04	-
		6.5. Passing the FP assessment does not depend on luck or supernatural powers.	59 (32.1%)	68 (37%)	61 (8.8%)	16 (8.8%)	22 (12.1%)	2.32	-
	political	6.6. Being taught and assessed in English creates more employment opportunities for me.	91 (49.5%)	70 (38%)	5 (2.7%)	11 (6%)	5 (2.7%)	1.73	-
		6.7. Being taught and assessed in English makes Oman an active part of the global village.	67 (36.4%)	84 (45.7%)	19 (10.3%)	8 (4.3%)	4 (2.2%)	1.89	-

As explained in Section 4.4.2.2, the questionnaire items were organised into groups for later analysis based on their themes. In some cases, shown in the table, this entailed recoding (i.e., changing responses to be 1=5, 2=4, 3=3, 4=2 and 5=1) so that semantically opposite or near-opposite items could be more directly compared. The expectation was not that the items in a group would be found almost totally equivalent, but that they might reveal broad trends of satisfaction, dissatisfaction and perhaps other feelings and perceptions, whilst anomalous response patterns might offer further insights.

These expectations were only partially met. In the large section on *Perceived Validity*, mean scores were fairly similar within and even between sub-sections: all were below 2.9 (3.0 =*No Opinion*), indicating broad but not overwhelming acceptance of FP assessment validity. In some other sections, however, means varied more widely and fell on both sides of the middle point (M=3.0). Some of these cases will be discussed below.

6.2.3. Means and Standard Deviations of Students' Responses to the Topics

This section aggregates the average responses to the individual items and presents the Mean and Standard Deviation of the responses to each topic to obtain an overview of students' perceptions of *Perceived Reliability*, *Perceived Validity, Preference of Tests*, *Satisfaction with Current Assessment Practices,* and *Impact* of *FP assessment*. Table 6.3 lists the questionnaire's topics which are hierarchically ordered according the means of the responses to each topic.

Table 6.3. Means of the Student Questionnaire's Topics

Topics	**Mean**	**Std. Deviation**
Political Impact	1.81	.78
Preference of Continuous Assessment	1.99	.75
Perceived Construct Validity	2.06	.75
Preference of Tests	2.36	.79
Perceived Content Validity	2.73	1.04
Perceived Reliability	2.83	.78
Social Impact	2.85	.58
Satisfaction with Current Assessment Practices	3.17	.51

The results showed that the students seemed to positively perceive the reliability and validity of the FP assessment as the mean scores were *Perceived Reliability* (M =2.83), *Perceived Construct Validity* (M=2.06) and *Perceived Content Validity* (M=2.73). It can be seen from the table that the students seemed to show less satisfaction with the content validity of FP assessment than they did with its construct validity. A closer look at the elements of the *Content Validity* topic reveals that the mean of one of its items was close to the disagreement range (i.e., $M \geq 3.1$). The means for the four items were respectively item 1.1: M = 2.7, item 1.2: M = 2.4, item 1.3: M = 2.9, and item1.4: M = 2.3 (see Table 6.2). The students' responses to the third item implied that their certainty level of how their achievement would be exactly assessed in the FP courses was not high. Actually, 41% of the students responded with *Disagree* or *Strongly Disagree* to this item, while 47% of the students responded with *Agree* or *Strongly Agree*. It seems that a considerable percentage of the students were ill-informed about how they would be assessed in their English language courses. The students' lack of knowledge about the assessment procedures could have lowered the average mean of FP *Perceived Content Validity*.

Another interesting point in Table 6.3 is that the mean score of the perceived *Political Impact* of FP was lower than that of perceived *Social Impact* of FP assessment. The means of the responses to FP *Political Impact* and *Social Impact* topics are 1.8 and 2.85 respectively, both of which fall in the agreement range (from M= 1 to M= 2.9). There seemed to be a majority agreement with the statements that indicates that FP assessment could entail considerable political impact by affecting the job opportunities in the labour market and the country's international status. Also, there seemed to be a moderate agreement with the idea that assessment in FP did not entail negative or drastic social impact on students' lives. Though FP assessment could be considered high-stakes, most of the students felt that the assessment was relatively fair, not frightening, and did not depend on luck (see Table 6.2). It is worth noting that though most of the students felt that both Continuous Assessment (CA) and tests were not stressful, they seemed to believe that the tests

(item 6.1: M= 2.26) were less stressful than the CA (item 6.2: M=2.1). Some of the reasons for this view were discussed in focus groups.

The topic of *Satisfaction with Current Assessment Practices* had the second highest mean (M= 3.17). This implies that most of the students were not satisfied with the FP assessment. Investigating the items under this topic shows that though most of the students seemed generally satisfied with the FP assessment (item 5.1: M=2.15), most of them also believed that the FP assessment should be changed (item 5.2: M= 3.42) and that the change should not include fewer assessment instruments (item 5.2: M=2.29). Interestingly, this response was found to conform to what most of the students said in the focus groups about increasing the number of assessment instruments (see Section 7.2.2.). In general, it could be concluded that the students seemed to be satisfied with the assessment practices, but they tended to believe that there should be more assessment instruments.

The last point about the means of the responses to questionnaire topics is that the respondents seemed to prefer AES continuous assessment (item 4.2: M= 2.23) more than the GES tests (item 3.1: M=2.46) (see Table 6.2). This preference resonated with their opinions as expressed in the focus groups as discussed in Section 7.2.3.1.

8.2.4. Comparing Perceptions amongst the Groups

This section further explores the students' responses to the questionnaire by investigating significant differences in responses to the items of each topic amongst the groupings by gender, college, specialization and self-evaluation. This exploration aims at identifying any clear pattern of consistent differences in the groups' responses that might shed some light on the participants' perceptions using filters such as college or gender.

6.2.4.1. Investigating Significant Differences Using Mann-Whitney U Test and Kruskal-Wallis Test

Before looking at the results of these tests it is important to clarify the rationale for using non-parametric tests. First of all, the data generated by a likert scale is an ordinal and sometimes categorical type of data that is best investigated using non-parametric techniques (Pallant, 2007). Second, these two tests were selected because the data set was found to be not normally distributed, so non-parametric tests are ideal in this situation (Fielding & Gilbert, 2006). The data set of students' responses was tested for normality of distribution using Kolmogorov-Smirnov test, skewness values, histograms and box plots. The results showed that the distributions of students' responses to each topic violated the assumptions of a normal distribution (see appendices 6.1, and 6.2). Thus, Mann-Whitney U Test was used to investigate significant differences between two groups and Kruskal Wallis Test was used amongst three groups or more.

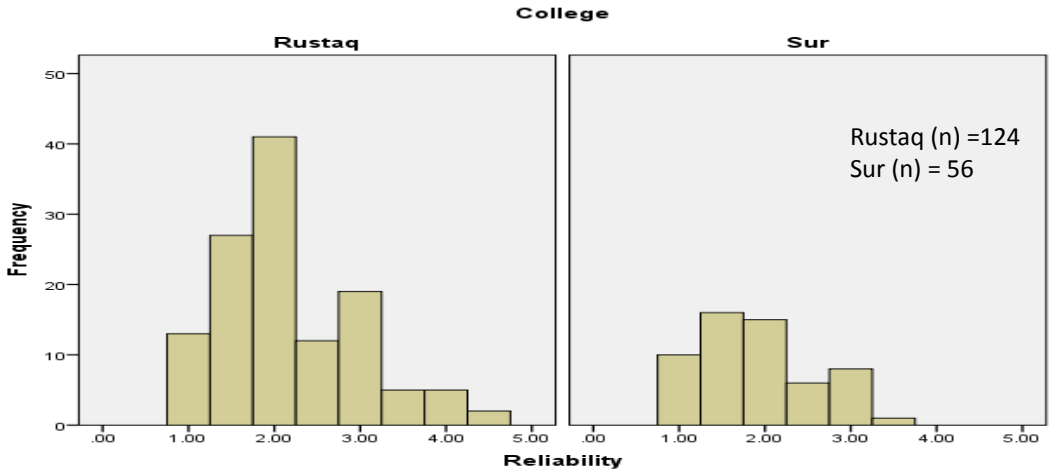
6.2.4.2. Differences between College Groups

In the college groups, a significant difference was found between Sur College's students (Mean Rank= 95.49, n= 56) and Rustaq College's students (Mean Rank= 79.45, n= 124) in their responses to *Perceived Reliability*, U= 2,853, $p<.05$. This indicates that the students at Sur College were significantly different in their perception of *Perceived Reliability* than the students at Rustaq College (see Figure 6.1). Sur Students viewed the assessment tools as more reliable than did their peers in Rustaq College. This can be deduced from the Mean Rank values presented above and Mean values in Table 6.4.

Table 6.4. Means of Students' Responses to Questionnaire Topics by Colleges

Topics	**Rustaq College's Students**		**Sur College's Students**	
	Mean	***n***	**Mean**	***n***
Social Impact	2.38	96	2.26	55
Preference of CA	1.93	125	2.01	55
Preference of Test	2.31	126	2.45	56
Political Impact	1.79	124	1.80	57
Perceived Reliability	2.16	124	1.90	56
Perceived Construct Validity	2.14	122	2.02	55
Perceived Content Validity	2.70	121	2.38	57
Satisfaction with Current Assessment Practices	2.83	114	2.71	52

Figure 6.1. Students' Reponses to *Perceived Reliability* by Colleges



6.2.4.3. Differences between Genders Groups

The results showed that the male and female students differed in responding to two topics namely *Preference of CA* and *Political Impact*. The Mann-Whitney U test revealed a significant difference between the female students' responses (Mean Rank=80.61, n=116) and male students' responses (Mean Rank= 106.13, n=62), U= 9,351, Z= -3.2, p=0.001). The female students showed more *Preference of CA* than did the male students. Likewise, the male students' responses (Mean Rank=102.52, n=62) and female students' responses (Mean Rank=84.18, n=118) were significantly different on the *Political Impact* of FP assessment: the female students seemed to believe that FP assessment had higher *Political Impact* more than did the male students, U= 2912, Z=-2.3, p=0.21. All in all, this means that the female students seemed to prefer CA more than did the male students and they seemed to emphasise the political impact of FP more than did the male students (see Table 6.5 & Figures 6.2 & 6.3).

Table 6.5. Means of Students Responses to Questionnaire Topics with Gender

Topics	**Male Students**		**Female Students**	
	Mean	*n*	*Mean*	*n*
Social Impact	2.32	58	2.34	92
Preference of CA	2.23	62	1.80	116
Preference of Test	2.45	62	2.30	118
Political Impact	2.00	62	1.69	118
Perceived Reliability	2.13	62	2.04	117
Perceived Construct Validity	2.09	61	2.11	115
Perceived Content Validity	2.57	61	2.59	115
Satisfaction with Current FP Assessment Practices	2.73	51	2.83	105

Figure 6.2.Responses to *Preference of CA* by Gender

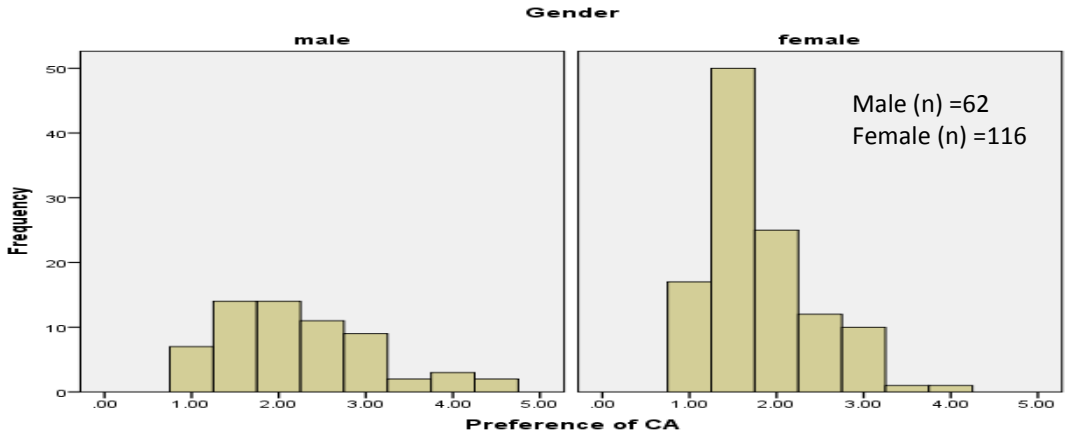
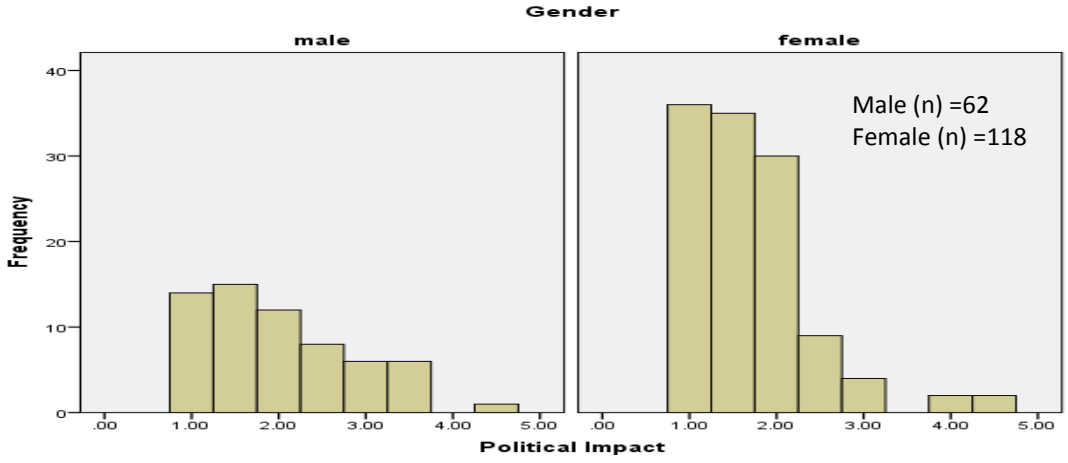


Figure 6.3. Responses to *Political Impact* by Gender



However, the difference between the female and male students' *Preference of CA* was not matched by a significant difference in their AES continuous assessment scores as shown by the results obtained from using Mann-Whitney U test. Actually, the female and male students' mean grades in both of the FP assessment instruments

(i.e., AES continuous assessment and GES tests) showed no significant differences (see Table 6.6). It can be seen from the table below that the mean grade for the female students is in general slightly higher than the mean grade of the male students in both courses. Intriguingly, when the distribution of the AES grades for each gender was examined, it was found that the female students performed better than the male students in the first quartile (i.e., the lower 25% of the AES grades distribution), but both genders obtained equal grades at the second quartile and the male students performed better at the third quartile (i.e., the higher 25% of the distribution, see Table 6.6). In the GES tests, the mean grades of the female students were higher in the first and second quartiles than those of the male students, but they become equal at the third quartile. This means that though the female students preferred CA more than did the male students, the female students only performed better at the lower end of the AES grade distribution, while the male students performed better at the higher end of the same distribution, and they obtained equal grades in the middle quartile.

Table 6.6. Mean and Quartile of Scores in GES and AES Courses by Gender

Gender			**GES**	**AES**
Male	*n*	Valid	61	61
		Missing	3	3
	Mean		1.82	2.71
	Quartile	25	1.30	1.70
		50	1.70	3.00
		75	2.30	3.70
Female	*n*	Valid	101	101
		Missing	18	18
	Mean		1.95	2.85
	Quartile	25	1.70	2.70
		50	2.00	3.00
		75	2.30	3.30

6.2.4.4. Differences among Self-Evaluation and Specialization Groups

In the first section of the questionnaire the students were asked to self-evaluate their language proficiency by selecting one of five levels to represent their language proficiency levels (i.e., 1= poor, 2=average, 3= good, 4=good, and 5= excellent). Kruskal Wallis was used to explore the differences, and the results showed no

significant differences in the students' responses to the questionnaire topics with regard to their self-evaluation groups. Similarly there were no significant differences in responding to the questionnaire's topics among the specialization groups.

To summarise, it could be concluded that the responses of participants to the questionnaire's items were not significantly different between the specialization and self-evaluation groups, but they were between the college and gender groups. The college groups showed a significant difference in responding to *Assessment Reliability*, and the gender groups showed a significant difference in responding to both *Preference of CA* and *Political Impact* topics.

6.3. The Teacher Questionnaire in Phase 1

6.3.1 Demographic Characteristics of the Participants

In the first phase, 27 teachers participated in responding to the questionnaire. Of these teachers, 10 were from Rustaq College (37%) and 17 were from Sur College (63%). In this sample, there were 17 male teachers (63%) and 10 female teachers (37%). There were five Omani teachers (18.5%) and 22 non-Omani teachers (81.5%). The participants' groupings by age and education are shown in the following tables.

Table 6.7. Number of Teachers in Age and Education Groups in Phase 1

Age	***n***	**%**	**Education**	***n***	**%**
20-30	9	33.3	**Diploma**	5	18.5
31-40	9	33.3	**BA**	15	55.6
41-50	3	11.1	**MA**	5	18.5
51-60	4	14.8	**PhD**	1	3.7
60+	2	7.4	**other**	1	3.7
Total	27	100.0	**Total**	27	100.0

6.3.2. Teachers' Responses to the Individual Items of the Questionnaire

A 30-items likert scale questionnaire was distributed to the teachers in the first phase of the study. The items were organised into six main topics, *Perceived Reliability, Perceived Validity, Tests vs. Continuous Assessment, Centrality in Assessment Writing, Experience in Assessment Writing* and *Impact*. The teachers' responses to

the items and the means of these responses are displayed in Table 6.8. This display of responses to each item in the questionnaire gives a detailed picture of the teachers' opinions and assists in explaining and understanding the means of each of the questionnaire's topics in subsequent sections. The means and recoded means of the responses to each item and topic were calculated in a similar way to that previously explained above in Section 6.2.1.

One of the interesting points that emerged when the responses to questionnaire items were compared was that, within one topic, the teachers sometimes agreed with certain items but not with others. The very different responses to some items which on the surface seemed similar conveyed a slightly more detailed view of the teachers' perceptions. For example, in the *Perceived Content Validity* topic, 74.1% of the teachers agreed with the statement that the assessment instruments represented the language skills of the curriculum, and 47.1% of the teachers agreed with the statement that the assessment instruments represented the objectives of FP courses. However, 40.4% of the teachers disagreed with the view that the assessment scores distribution on language skills reflected time spent on teaching these skills in classroom. This example reveals that the teachers agree that some aspects of content validity are represented in FP assessment practices but not others. Similar examples are discussed in the next section.

Table 6.8. Frequency and Means of Responses to Teacher Questionnaire Items in Phase 1

Topic	Sub-Topic	Items	SA 1	A 2	NO 3	D 4	SD 5	Mean	Mean if Recoded
Perceived Reliability	_	1.1. The criteria and the rating scales that the students are assessed by in FP facilitate assessing them consistently.	1 (3.7%)	11 (40.7%)	7 (25.9%)	7 (25.9%)	-	2.77	_
		1.2. The assessment instruments in FP are consistent in evaluating the students' language performance.	1 (3.7%)	1 (29.6%)	8 (29.6%)	8 (29.6%)	1 (3.7%)	3	_
		1.3. I am satisfied about the reliability of the assessment instruments implemented in FP.	-	12 (44.4%)	7 (25.9%)	8 (29.6%)	-	2.85	_
Perceived Validity	content	2.1. The scores on the different assessment instruments reflect the time spent on teaching the English language skills.	1 (3.7%)	8 (29.6%)	4 (14.8%)	9 (33.3%)	2 (7.4%)	3.13	_
		2.2. The assessment instruments in FP represent the English language skills and activities covered in the curriculum appropriately.	3 (11.1%)	17 (63%)	1 (3.7)	6 (22.2%)	-	2.37	_
		2.3. The assessment instruments represent efficiently the courses objectives in FP courses.	1 (3.7%)	12 (44.4%)	5 (18.5%)	8 (29.6%)	-	2.77	_
	predictive	2.4. The assessment instruments implemented in FP inform on students' abilities to linguistically handle the academic courses in the First Year.	2 (7.4%)	13 (48.1%)	5 (18.5%)	5 (18.5%)	-	2.52	_
		2.5. The FP English assessment prepares students well to cope with the language demands of their academic courses.	1 (3.7%)	9 (33.3%)	6 (22.2%)	8 (29.6%)	3 (11.1%)	3.11	_
	inappropriateness	2.6. The assessment instruments used in my courses are appropriate in assessing students' English language abilities. (Recode)	-	13 (48.1%)	6 (22.2%)	7 (25.9%)	1 (3.7%)	2.81	3.15
		2.7. The FP assessment instruments should be changed.	1 (3.7%)	7 (25.9%)	5 (18.5%)	10 (37%)	3 (11.1%)	3.27	_
		2.8. There should be less assessment instruments (continuous assessment and tests) in FP courses than what is there currently.	3 (11.1%)	8 (29.%)	6 (22.2%)	9 (33.3%)	1 (3.7%)	2.37	_

	construct	2.9. The current assessment instruments are valid.	1 (3.7%)	11 (40.7%)	9 (33.3%)	5 (18.5%)	1 (3.7%)	2.78	_
		2.10. The FP assessment instruments provide teachers with suitable information about their students' English language performance.	1 (3.7%)	10 (37%)	7 (25.9%)	8 (29.6%)	1 (3.7%)	2.93	_
		2.11. The students' scores in FP assessment instruments represent their language performance levels accurately.	1 (3.7%)	9 (33.3%)	7 (25.9%)	9 (33.3%)	-	2.92	_
Test/CA		3.1. Tests are more valid than continuous assessment.	3 (11.1%)	6 (22.2%)	3 (11.1%)	13 (48.1%)	2 (7.4%)	3.23	_
		3.2. Tests are more reliable than continuous assessment	3 (11.1%)	5 (18.5%)	2 (7.4%)	15 (55.6%)	2 (3.7%)	3.19	_
Centrality in writing assessment		4.1. Teachers should write their own final tests locally at the colleges. (Recode)	2 (7.4%)	12 (44.4%)	4 (14.8%)	6 (22.2%)	3 (11.1%)	2.85	3.14
		4.2. Teachers should write their own continuous assessment locally at the colleges. (Recode)	7 (25.9%)	9 (33.3%)	3 (11.1%)	7 (25.9%)	1 (3.7%)	2.48	3.52
		4.3. Teachers should conduct the same assessment instruments in all of the six colleges.	1 (3.7%)	6 (22.2%)	6 (22.2%)	13 (48.1%)	1 (3.7%)	3.51	_
Confidence in marking/writing		5.1. I am confident about my ability to write final tests and assessment activities for FP students.	10 (37%)	7 (25.9%)	7 (25.9%)	3 (11.1%)	-	2.11	_
		5.2. I need more training to write reliable and valid assessment instruments. (Recode)	2 (7.4%)	15 (55.6%)	4 (14.8%)	2 (7.4%)	2 (14.8%)	2.66	3.33
		5.3. I have appropriate experience in marking tests and assessment tasks using provided scales.	7 (25.9%)	16 (59.3%)	1 (3.7%)	3 (11.1%)	-	2	_
Impact	social	6.1. I have made the students aware of the consequence of failing/passing FP assessment.	2 (7.4%)	4 (14.8%)	11 (40.7%)	7 (25.9%)	2 (7.4%)	3.12	_
		6.2. As far as I know, the department is taking sufficient account of the probable social consequences of failure in the FP assessment to students.(e.g. making students and teachers aware of those consequences, working to avoid the severity of the consequences)	7 (25.9%)	9 (33.3%)	6 (22.2%)	3 (11.1%)	-	2.2	_

		6.3. The assessment instruments are fair enough to students that they should be carried out in the same way in future.	3 (11.1%)	7 (25.9%)	9 (33.3%)	6 (22.2%)	2 (7.4%)	2.89	_
		6.4. I have the opportunity to give feedback on the quality of the assessment instruments.	2 (7.4%)	7 (25.9%)	9 (33.3%)	9 (33.3%)	-	2.92	_
		6.5. Other parties (students, society, researchers and other organizations) have the opportunity to give feedback on the quality of the assessment activities and tests.	3 (11.1%)	7 (25.9%)	6 (22.2%)	9 (33.3%)	2 (7.4%)	2.93	_
	political	6.6. The Omani National Standards for the Foundation Programme and the foundation programme audit are vital to ensure accountability in English language teaching institutions.	1 (3.7%)	14 (51.9%)	10 (37%)	1 (3.7)	1 (3.7%)	2.52	_
		6.7. Assessing the academic courses in English helps to develop the country's economy.	4 (14.8%)	13 (48.1%)	6 (22.2%)	3 (11.1%)	1 (3.7%)	2.41	_
		6.8. I think those students' scores in FP English language assessment should not be a gate-keeper to higher education in Oman. (Recode)	5 (18.5%)	7 (25.9%)	10 (37%)	2 (7.4%)	3 (11.1%)	2.66	3.33

6.3.3. Means and Standard Deviations of Teacher Responses to Questionnaire

The topics of the questionnaires that cover some aspects of FP assessment are displayed in an ascending order based on their means in Table 6.9. It should be remembered that lower means indicate agreement, higher means indicate disagreement, and a mean of 3.0 indicates equal strength of agreement and disagreement with a certain item. The following paragraphs discuss the topics with the highest, lowest and middle means.

Table 6.9. Mean and Standard Deviation for Teacher Questionnaire Topics in Phase 1

Topics	**Mean**	**Std. Deviation**
Confidence in Writing and Marking	2.48	.81
Political Impact	2.75	.48
Content Validity	2.76	.78
Perceived Predictive Validity	2.84	.87
Perceived Construct Validity	2.84	.75
Social Impact	2.89	.64
Perceived Reliability	2.92	.77
Inappropriateness	2.93	.49
Test /CA	3.19	1.11
Centrality of Assessment Writing	3.37	.99

The teachers' *Confidence in Writing* and *Marking Assessment* in FP had the lowest mean M=2.48. This indicates that the teachers seemed to believe that they were able to handle writing and marking FP assessment. Investigating the mean for each item under this topic showed that the responses to items 5.1 and 5.3 in this topic were within the *Agree* range ($1 \leq M \leq 2.9$). This implies that there was large agreement with the items on teachers' confidence about their skills in writing and marking FP assessment; however the mean for item 5.3 was M=3.33 indicating a need for additional training (see Table 6.8 for the responses to each item).

The topic with highest mean was *Centrality in Assessment Writing* M=3.37. A closer examination of the means for the individual items that comprised this topic suggested

a mild disagreement and revealed that 51.8% of the teachers seemed to believe that test should be conducted locally at the colleges; and 59.2% of them disagreed with the view that continuous assessment should be centrally controlled. Similarly, most of them seemed to believe that assessment instruments did not need be the same in all of the six colleges.

Most teachers seemed to feel that English language assessment did not have any strong negative social impact. The Mean value of the teachers response to the *Social Impact* topic is 2.89 indicating weak agreement with the items under this topic; most of the teachers agreed that FP assessment was fair, stakeholders had the opportunity to give feedback, and the FP English department was aware of and worked to minimise the negative social consequences of FP assessment. From the same table, it can be seen that a high percentage of the teachers selected the 'No Opinion' option in all of the *Social Impact* items. This should be considered when making future recommendations and should be validated through the triangulation of the findings produced by the other data collection methods.

Between the highest and lowest means of the teachers' responses to the questionnaire's topics fall the means of responses to the FP assessment validity subtopics: *Perceived Content Validity* (M=2.76), *Perceived Predictive Validity* (M=2.84), *Perceived Construct Validity* (M=2.84) and *inappropriateness* (M=2.93). All of these values, except the last one, implied a seemingly moderate satisfaction with the validity of FP assessment. *Content*, *Predictive* and *Construct Validity* of FP assessment were rated similarly. However, the teachers responses to the items in *inappropriateness* suggest that in general FP assessment was appropriate for its purposes and should not be changed, but the assessment instruments should be decreased. The last view contradicts the teachers' views on increasing the number of assessment tasks as expressed in the interviews (see Section 7.3.4.2). In general the teachers' responses on FP assessment validity seem to imply a general satisfaction with the content, construct and predictive validity of FP assessment, but indicate that FP assessment does not prepare students for upcoming academic study (item 2.1), time spent on teaching specific skills is not reflected in assessment (item 2.5), and

the number of assessment instruments should be decreased (item 2.8). This repeated pattern of responding differently to some items with one broad aspect of FP assessment reveals the complexity of teachers' perceptions, and confirms that, as anticipated at the research design stage, the best way to understand the students' and teachers' perceptions is by investigating their responses to each item as well as considering the mean responses to the general topics.

6.3.4. Investigating Significant Differences in Teachers' responses among the Groups

This section investigates the significant differences in teachers' responses to the questionnaire's topics among the groupings by age, gender, nationality, and college. The teachers' responses to each of the topics and subtopics were tested for significant differences using Mann-Whitney U test between two groups and a Kruskal-Wallis test among more than two groups.

As has been explained in Section 6.2.4 above, these tests were employed because the type of data generated by likert scales is usually ordinal and sometimes categorical, and such tests are ideal for this type of data. Also, tests of normality of distribution have shown that the distributions of data were skewed (see appendices 6.3 and 6.4).

6.3.4.1. Differences between Gender Groups

In the gender groups, though there was a considerable difference between the means of the female and male teachers' responses to the items in the *Perceived Construct Validity* topic, but no significant differences were found.

6.3.4.2. Differences between Nationality Groups

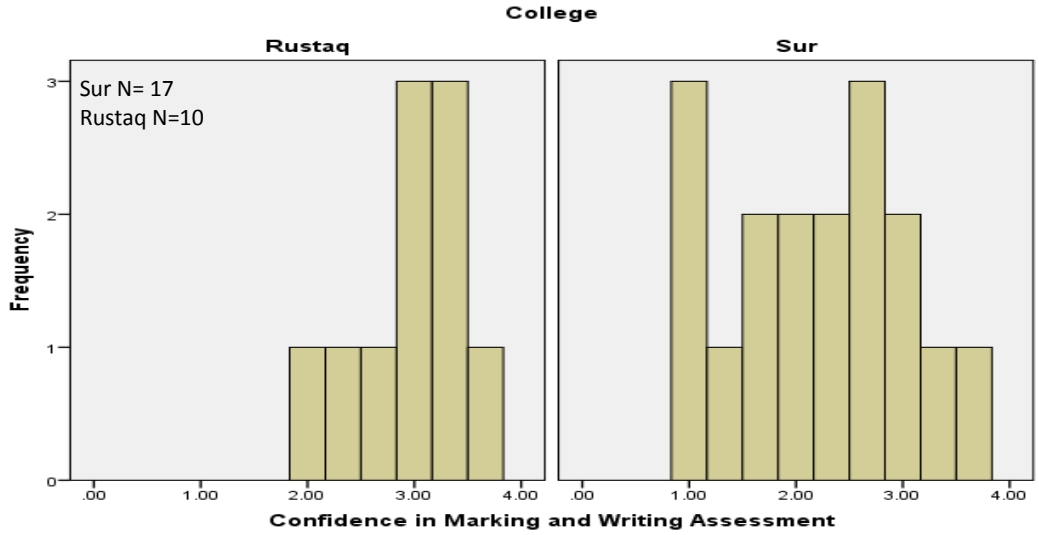
The differences in means of responses to the questionnaire topics between the nationality groups were also investigated. The Omani and non-Omani groups differed in responding to the *Test/CA* and *Confidence in Writing and Marking Assessment* topics. The significance of this difference was tested using Mann-Whitney U test which revealed no significant difference between the Omani (Md = 4,

n=5) and non-Omani (Md = 3.5, n=21) groups in their responses to the *Test/CA* topic U = 42, z = -.71, p =.47.

6.3.4.3. Differences between College Groups

However, a significant difference was found in the teachers' responses to the items of *Confidence in Marking and Writing assessment* between the college groups, Sur (M=2.1) and Rustaq (M= 2.9). Using Mann-Whitney U test to compare teachers' responses in Sur College (n=17, Mean Rank=11.21), and Rustaq College (n=10, Mean Rank= 18.75), the results showed a significant difference between the groups, U = 37.5, Z= -2.4, p < 0.5. The Sur College teachers seemed to be more confident about their assessment writing and marking skills than were the Rustaq College teachers.

Figure 6.4. Mann-Whitney Results of Confidence in Marking and Writing Assessment by Colleges



6.3.4.4. Differences among Age Groups

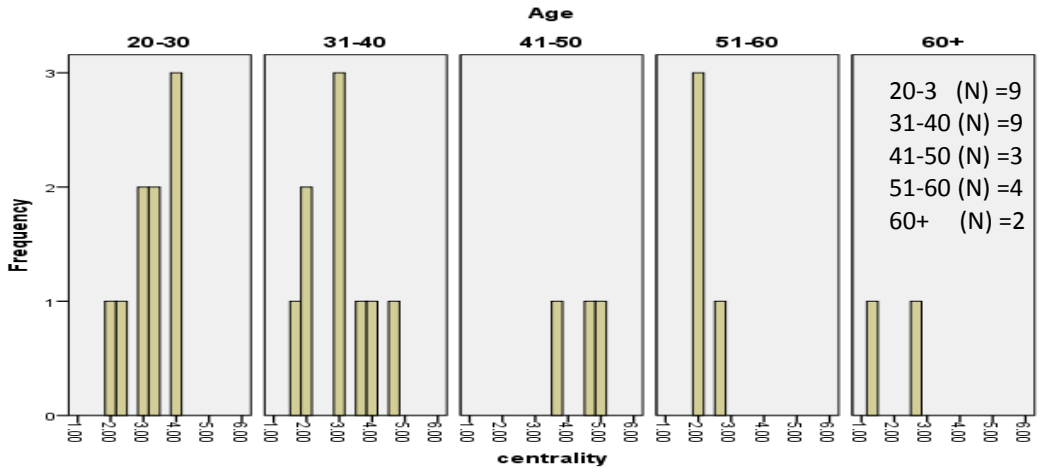
The other significant difference was found among the age groupings by teachers in responding to *Centrality of Assessment Writing* topic (Table 6.12). Two age groups (i.e., 20-30 years old, and 41-50 years old) seemed to disagree with controlling and writing FP assessment centrally while two other age groups (i.e., 51-60 years old and 60+ years old) seemed to agree with it. The age group 31-40 years old did not express a clear opinion on this issue (M= 3.0). Kruskal –Wallis test was used to

evaluate the difference between the age groups, (Gp1, n= 9:20-30), (Gp2, n=9: 31-40), (Gp3, n=3, 41-50), (Gp4, n= 4:51-60), (Gp5, n=4: 60+), in their responses to the preference of *Centrality in Assessment*. The results showed a significant difference (4, N=27) = 11.15, p = .025 (see Figure 6.5), meaning that centrality of FP assessment was viewed differently by the teachers according to their age groups, and that centrality in assessment was preferred by the two older age groups and opposed by two younger age groups. These results should, however, be considered very cautiously since three of the five groups included less than seven participants which is the lowest number acceptable for the Kruskal-Wallis Test. Any further discussion of this finding will be based on the descriptive statistics only not on the inferential ones.

Table 6.10. Means of Reponses to Preference of Centrality in Age Groups

Age Group	**Mean**	**N**	**Std. Deviation**
20-30	3.22	9	.726
31-40	3.00	9	1.00
41-50	4.44	3	.69
51-60	2.16	4	.33
60+	2.00	2	.94
Total	3.03	27	.99

Figure 6.5: Kruskal-Wallis Test of Age Groups with Responses to Centrality of Assessment



6.4. Discussion

The present section compares the students' and teachers' perceptions of (1) FP assessment validity and reliability, (2) the social and political impact of FP

assessment, and (3) tests versus CA. This is followed by a further discussion of the teachers' views on centrally controlled assessment.

Before discussing the results in these four main areas, it should be noted that the items with similar topics in the student and teacher questionnaires were not always identical. The items were designed to address issues most relevant to students in the student questionnaire and to the teachers in the teacher questionnaire. However, the discussion of the results focuses on the general results of the questionnaires' topics, not their individual items, and there are common areas between the questionnaires that can be compared.

The table below compares the mean responses of the student questionnaire topics to the mean responses of the teacher questionnaire topics presented in an ascending order. As mentioned earlier, the mean (M=2.9) and lower signifies agreement with the topic, while the mean (M= 3.1) or more signifies disagreement.

Table 6.11. Comparing Means of Student and Teacher questionnaires

Student Questionnaire Topics	**Mean**	**Teacher Questionnaire Topics**	**Mean**
Political Impact	1.81	Confidence in Writing and Marking	2.48
Preference of CA	1.99	Political Impact	2.75
Perceived Construct Validity	2.06	Content Validity	2.76
Preference of Tests	2.36	Predictive Validity	2.84
Perceived Content Validity	2.73	Construct Validity	2.84
Perceived Reliability	2.83	Social Impact	2.89
Social Impact	2.85	Reliability	2.92
Satisfaction with Current Assessment Practices	3.17	Inappropriateness	2.93
		Test /CA	3.19
		Centrality of Assessment Writing	3.37

6.4.1. Teacher Perceptions of FP Assessment

In general, both student and teacher responses to the questionnaires tended to reflect a positive perception of the validity and reliability of the assessment instruments but a slight dissatisfaction of the assessment instruments implementation. Both the students and teachers perceived the assessment instruments validity more positively than their reliability as shown by the lower means for validity in Table 6.11. However, the students and teachers seemed to differ in their satisfaction levels with

FP assessment. The students' responses to the topic on *Satisfaction with Current Assessment Practices* indicated dissatisfaction (M=3.17), while the teachers responses to *Inappropriateness* of FP assessment indicated moderate satisfaction (M=2.93).

6.4.2. Student and Teacher Views of FP Assessment Impact

Interestingly, teachers' and students' responses to the *Social and Political Impact* topics were similar in rating the political impact higher than the social impact and in implying that FP assessment entailed political but not serious social consequences on students, teachers and society. They seemed to recognise the "prestige" and importance of the English language assessment for future national employment and international status of the country, as well as its role as a gatekeeper to higher education; this finding conforms to previous studies (e.g., Shohamy, 1996; Ross, 2008). However, both the teachers and students disagreed with the view that FP assessment had drastic social consequences. The teachers' responses to the items about the procedures for minimizing the negative social impact of assessment showed that most of them seemed to agree that FP assessment did not imposed negative social consequences. Equally, the social impact was not considered great in the students' responses. Their responses showed a majority agreement with the items that suggested that FP assessment was fair, not frightening and not stressful. This finding is substantiated by the findings obtained from the focus groups and interviews in Chapter 7.

When the students were asked whether the tests and CA were not stressful, their responses were positive. But they agreed more with the statement "tests are not stressful" than with the statement "CA is not stressful". This finding is in line with the argument that performance based tasks involve communication stress or anxiety which may well influence students' performance along with other factors (Bachman, 2002; Phillips, 2011). It was also found that performance assessment did not produce better results than test in terms of the writing skill (Hamp-Lyons, 1997); therefore, if AES assessment, which includes performance based tasks, against common expectations (e.g., of teachers), did not provide a less stressful environment than

tests, and did not result in better performance, the advantages of using this type of assessment in FP assessment should be reviewed.

Furthermore, the results revealed that a significant difference in the responses on the political impact topic between the female and male students. Female students agreed more strongly with the statements on FP political impact than did the male students. No significant differences were found between the genders in teachers' responses to FP assessment political and social impact though. This, as will be discussed in Chapter 11, might be explained by the challenges female graduates face in the labour market.

6.4.3. Tests vs. CA in Student and Teacher Perceptions

The results showed that the teachers' and students' responses differed with regards to their preference for tests and/or CA. Most of the teachers in the sample thought that the tests were not more valid or reliable than CA. Most of the students preferred CA over the tests, although, both were considered useful.

These results differ from those of a recent study by Cheng, Andrews, and Yu (2011) which explored students' and parents' perceptions on the traditional examinations compared to a currently applied School Based Assessment (SBA) system in China. They found that no significant difference in how students viewed SBA and exams. They also reported that the students differed in how they percieved the SBA and exams items based on their self-reported language levels; "students with high perceived language competence responded more positively to the items relating to the external examinations while students with low perceived language competence responded more positively to the items relating to SBA" (p. 238). The results of the current study did not report any significant difference amongst students' self-evaluations of their language proficiency levels in their perception of CA compared to tests. Actually all self-evaluation groups responded more positively to CA than they did to the tests.

6.4.4. Centrality of Assessment Writing in Teacher Perceptions

The teachers' responses to the items on writing FP assessment centrally showed a significant difference among the age groups. The 20-30, and 41-50 age groups generally disagreed with using central assessment, while 51-60 and 60+ groups generally agreed with centrality in assessment. Research on centralization of assessment in schools (Runte, 1998) suggested that it could deskill teachers in the assessment domain, and teachers might feel threatened by it. The same study found that most teachers did not have adquate assessment skills, and their skills were picked up through apprenticeship not through proper training courses: thus being involved in centrally managed assessment assisted in enriching their skills.

Though using centralised tests as a gatekeeper to higher education is common in Asian countries (Ross, 2008), democratic assessment where the stakeholders are involvoed in the process of decision making to avoide misuses of the power of tests has been ecouraged (Shohamy, 2001). Nonetheless, there are voices that warn that using teacher based assessment rather than centralised tests does not remove the power of assessment; it just shifts the control from central bodies to teachers (Gipps, 1999). In this study, most of the teachers seemed to prefer centralization of tests but not centeralization of CA, though the majority of the teachers expressed their confidence of their assessment writing and marking skills; one of the reasons for this adherence to centralised tests could be that some teachers seemed to consider tests to be more strict and objective than CA, as will be further discussed in presenting the findings from the teacher interviews in Chapter 7.

6.5. Summary and Concluding Remarks

In this chapter, the findings obtained from a student questionnaire and teacher questionnaire conducted in the first phase of this study were presented. The questionnaires addressed a number of topics on FP assessment validity, reliability and impact; they also surveyed the participants' satisfaction with FP assessment and their perceptions of the two assessment instruments used (i.e., tests and CA).

The results indicated that generally FP assessment validity was positively viewed by both students and teachers. However, the students seemed to be dissatisfied with FP assessment while the teachers seemed to be moderately satisfied. Also, the female participants seemed to acknowledge the political impact of English language assessment more than the male participants. This view could be tentatively explained by the low employability rate of women compared to men in Oman as well as in other Gulf countries (Klasen & Lamanna, 2009), and by the extra challenges women face when attempting to attain middle to upper management positions (Al-Lamky, 2007). The female students' preference for CA over tests was found to be significantly stronger than the male. Furthermore, most teachers' responses to using centrally controlled assessment significantly differed according to the teachers' age groups; they seemed to prefer writing tests and CA locally at the colleges. However these views seemed to vary according teachers' age groups. Some of these findings recur in the coming chapters and their implications will be discussed in Chapter 11.

Chapter 7: Results from Student Focus Groups and Teacher Interviews in Phase 1

7.1. Introduction

This chapter includes three main sections: the student focus groups' results, the teacher interviews' results and a discussion. It starts with an outline of the study questions that are addressed by focus groups and interview results. Second, it identifies common themes that emerged from the student focus groups and presents them in three main sections: uncertainties about assessment, perceptions on General English Skills (GES) assessment, and perceptions on Academic English Skills (AES) assessment. Third, it presents the main themes that recurred in the teacher focus groups and divides them into three main sections similar to the ones used in presenting the focus groups results: uncertainties about assessment, perceptions on GES assessment, and perceptions on AES assessment. Finally it concisely discusses the results and links them to previous pertinent studies.

Box 7.1. The Study Questions addressed by the Focus Groups and Interviews Results

1.	How well did the process of assessing students' English language performance, through classroom assessment and tests, function in the Foundation Programme (FP)?
1.2.	How were the reliability, validity and effectiveness of FP assessment viewed by the students and teachers?
1.3.	How was the impact of FP assessment perceived by students and teachers?
1.4.	What were the differences between the 'continuous assessment' model used in the AES course and 'test' model used in the GES viewed in terms of effectiveness, accuracy, preference?

7.2. Student Focus Groups

This section presents the results attained from the student focus groups which are intended to partly answer the questions of the study stated above. The focus groups were conducted in two colleges namely Rustaq College and Sur College. Table 7.1 displays an overview of the college, gender, participant numbers and length of each focus group. Also, all of the 184 students who completed the questionnaire forms

were invited to take part in the focus groups and 106 of them agreed to participate. In this phase, 12 focus groups were conducted seven of which were female only groups and the other five were male only groups. As mentioned in Chapter 4, in the pilot study, the participants expressed their preference of gender specific focus groups.

Table 7.1. An Overview of the Participants in Phase 1 Focus Groups

Group	**College**	**Gender**	**Number of Students**	**Length/minutes**
Group 1	Rustaq	F	12	53 min.
Group 2	Rustaq	F	8	32 min.
Group 3	Rustaq	F	16	32 min.
Group 4	Rustaq	F	9	35 min.
Group 5	Sur	M	9	33 min.
Group 6	Rustaq	F	6	32 min.
Group 7	Sur	F	12	38 min.
Group 8	Rustaq	F	13	26 min
Group 9	Sur	M	8	38 min
Group 10	Sur	M	3	14 min.
Group 11	Sur	M	7	34 min.
Group 12	Rustaq	M	13	51 min.
Total			**106**	**418 min.**

All of the groups' discussions were carried out in Arabic and were video-taped. The recordings were translated into and transcribed in English. Though the transcriptions were produced as literal translations, incomplete phrases, repetitions, and non-linguistic communication signals were intentionally excluded. This deliberate omission in transcriptions was based on the fact that this study used thematic analysis which focused on what was said not how it was said (Bryman, 2004). Also, it was hoped to produce more comprehensible and coherent transcriptions by this exclusion. In the discussions, more than one student spoke simultaneously at several occasions, and the flow of the discussion became fragmented. Thus, to fully capture the students' opinions, the process of transcription focused on completing and following their expressed views by disregarding any trivial interruptions. Nonetheless, some interruptions were transcribed when they entailed comprehensible ideas. Therefore, translating and transcribing the discussions followed a less *literal translation* which will be referred to as *edited translation* for the purposes of this study. A sample transcription of the first page of the second focus group's discussions is presented in both literal translation and edited translation from Arabic in appendix 7.1. It could be

noticed from the appendix that the scripts are similar in terms of representing the students' views.

The transcripts were coded and the common themes were identified following the steps described in Section 4.7.2. Coding and analysing focus groups transcripts resulted in 20 codes that were categorised into three main themes: uncertainty about assessment instruments' weightings and scales, tests in students' perceptions, and continuous assessment in students' perceptions. The latter two themes were divided onto six subthemes discussed in the subsequent sections.

7.2.1. Uncertainties about GES and AES Assessment Instruments

The majority of the students in seven focus groups tended to express uncertainty about how the scores were distributed on the main FP assessment instruments. Though, all of them seemed to be aware that the assessment of the General English Skills (GES) course included a midterm test and a final test and the assessment of the Academic English skills (AES) course included essay writing and a presentation, many of the students seemed uncertain about the scores distribution. The following extracts reveal this uncertainty as demonstrated in the students' discussion of the weightings of the assessment instruments. This extract comes from Focus Group 9.

> Student 3: In the GES course, there will be 50% of the total scores on the presentation and 50% on the essay.
> Student 7: It is still not very clear how the scores in the AES course are divided. Some teachers say the project is allocated 50% of the total scores while other teachers say that it is worth 20% of it only, so we do not know yet…
> Student 3: It is not clear. In the GES course, we have a speaking test and in the AES course, we have a presentation. So are they accumulated and how many scores each is awarded?
> Student 6: we said that the vision is not clear in regard to scores distribution.
> Student9: The total score is 220 since the speaking interview is worth 20 scores, the mid and the final exams are 50 scores each. The AES course contains 50 scores for the essay and the other 50 scores are for the presentation. So it is really confusing.

The uncertainty about the weightings of the assessment instruments expressed in this focus groups resonated with the other focus groups. The following extract is an instance. Most of the students in this group generally seemed to believe that the

scores distribution of the GES assessment instruments was 50/50 whereas it was actually 40/60.

> Student 3: The exams should be given more scores; a combination of 60% to exams and 40% the types of assessment is better than what it is now 50/50.

In addition to the uncertainties about scores distributions, many students seemed ill-informed about the criteria used in marking scales to evaluate their language performances in the essay and presentation. When asked about how scores would be given, most of them were aware that their teachers would be using marking scales but seemed oblivious of the scales' criteria. Few students, in three groups only, mentioned several criteria of the scales such as: eye-contact, posture and grammar with regard to the presentation marking scale; and grammar, organization and content with regard to the essay marking scale. The following extracts, from three other focus groups, manifest the lack of clarity of the marking scales as experienced by many students.

> **Group (3)**
> Student 1: This semester the way we are going to be assessed is not clear. No one explained to us. The course plan seems not stable and the teachers seem not sure of how and what the assessment will look like.
>
> **Group (7)**
> Student 4: We do not know how we will be assessed in the essay, what we know is that the font should be Times New Roman, size12, and the lines should be double spaced. But she says nothing about grammar, organization or content.
>
> **Group (8)**
> Student 4: In our group we do not know how the essay or the presentation will be marked our friends from the other groups tell us that they know about how the scores are divided and they let us know.

Students' uncertainties were not limited to the distribution of scores and marking scale criteria, there were also uncertainties about the test sections. In one group, few students stated their confusion about whether grammar would be included in the midterm and final tests or not. In another group, the students' discussion suggested that they were not aware of the fact that grammatical rules were actually tested in a section in the test titled "Language Knowledge". They kept on speculating whether the grammar rules would be tested or not. Part of this discussion is presented below.

Group (4)

Student 1: The midterm test will include reading, writing, listening and grammar however; the flyer distributed to us does not mention grammar. It says something about language Knowledge.

Student 2: They did it last semester, they told us that grammar will be in the exam and then it was not there.

7.2.2. GES Tests in Students' Perceptions

The presentation of students' views of FP assessment categorises the views into two main sections: views about GES tests and views about AES assessment. In each of the categories, examples of students' discussions were provided to reflect the students' perceptions of validity aspects of assessment such as: content validity, construct validity, reliability and impact.

7.2.2.1. The Content of GES Tests

In the focus groups, issues about the content of the midterm test and forthcoming final test were raised and debated. One of the issues was the difficulty of the reading tasks in which they faced new topics. One instance mentioned was that in the midterm test the reading test task was about "human cannonball". They said it was a new type of sport that they had never heard of before, thus they found it difficult to respond to the task questions. Group (11) discussed this issue saying:

Student 3: The reading passage was incomprehensible without the picture.

Studen6: True, it was about an unknown kind of sport.

Student 1: We have never heard of such a thing so comprehending what the passage was about, was so difficult.

It was discussed that in the midterm test, it was not only the topics that were new to them, but also the vocabulary of the reading task and length of the writing task.

Group (11)

Student 1: The topics used in the exam should be related to the topics studied in class because the vocabulary used in the test should be similar to the vocabulary used in the course.

Student 2: The test includes many things that we have not learned in class.

The second issue raised was about the lack of proper preparation for the grammar test tasks. Though the midterm and the final tests allocate only 10% of the total mark to the Language Knowledge task (i.e., grammar and vocabulary test items, see Section 5.3.2 for a breakup of the test scores), almost in all focus groups, the students expressed their need for additional grammar tutorials. In several groups, it was reported that even though the textbooks included activities on grammar rules, the teachers tended not to teach them. This was because they, as many students believed, seemed to be unqualified to teach grammar rules or because the teachers expected the students to study the rules by themselves as a form of autonomous learning. The students felt that more grammar lessons were needed to succeed in their future academic study and to pass the GES tests as manifested throughout the following extracts.

Group (9)

Student 1: We did the midterm exam and it was very difficult. We had not been given any practice quizzes before it. It was a shock.

Student 2: Our teachers do not explain grammar and we found the grammar part of the test very difficult. None of the teachers discuss grammar with the students.

Group (7)

Student 2: There is little about grammar in the book but there is a lot about it in the exam.

Student 13: Teachers rarely discuss grammar and we need it.

Student 9: Teachers think that we do not need grammar and they teach us about complicated things assuming that we have learned grammar. We have not, we need to learn grammar and start from simpler levels.

The listening task, on the other hand, seemed to be considered by almost all groups as the most difficult task of the GES tests. Likewise, it was claimed that the listening activities undertaken in classroom were simpler than what was in the tests and fewer than what was needed to be able to perform well in the listening test tasks. The difficulty of the listening test task was conveyed in the extracts below.

Group (11)

Student 2: The listening (part two) was so difficult. We are not used to such a thing. We need a book on listening to practice listening.

Student 5: It was very quick; we could not answer the questions in the pace that we were supposed to.

Group (8)

Student 6: We did the midterm test and it was very difficult we were not given any quizzes or practice. It was a shock.

Student 8: We will demonstrate against it.

Student 1 3: The listening part of the exam was the most difficult one. We could not hear what was on the tape.

7.2.2.2. What the GES Tests Assess

In the focus groups, concerns were raised about what GES tests assessed. Some of these concerns were about limited test time, difficulty levels of the test tasks and whether the tests really assessed English language skills or not. The majority of the students seemed to believe that the short test time negatively affected what the tests assessed; they asserted that with time constraints, they could not show their actual language abilities through the tests as explicated in the comments below.

Group (1)

Student5: No, tests are not enough to show the language levels of the students. You have to answer 6 to 7 pages of questions in a very limited time. This cannot be an accurate measure of students' language levels ... Also, you have two long texts to read and respond to their subsequent questions but time is limited to two hours. How could you do that?

Group (8)

Student 3: We generally like the assessment tools, we learn from them. Even the midterm test, shows us how the final will look like and gives us some practice for it. But we do not learn from the final exam.

Student 5: In the exam, we usually do not know what to do, the time is limited. We need more time to understand the questions, we cannot read quickly.

In the two groups below, a few students suggested that the test tasks aim at a high level of proficiency in language and that they did not discriminate between the students' language proficiency levels. It was indicated that the difficulty of test items hindered some students from attempting to answer them. The students also seemed to believe that different students performed differently in the tests and CA regardless of how much effort they put into it; some were better at the tests while others were better at CA. In group 8, three students talked about how each one of them performed better in a certain type of assessment instrument.

Group (8)

Student 6: … I usually do not participate a lot in class and do not study well for the tests but my scores are always surprisingly high. I do not cheat, but I believe that the language depends on individuals natural skills.

Student 2: On contrary, I participate a lot in the classroom. And I study hard but in the tests I do not get the scores I deserve. The presentation and the essay's scores represent my skills more.

Despite the expressed difficulty of some test tasks, many students seemed concerned about not being able to cope with the FY language requirements in the coming semester. The effectiveness of the FP assessment to distinguish linguistically ready students for FY study from linguistically unready students appeared to be repeatedly questioned in the focus groups.

Group (11)

Student 4: We know we will pass the foundation, but we will face a difficulty the coming year most first year students are struggling in the first year courses and their GPAs are low, some of them willingly dropped their courses and went to find jobs elsewhere.

Student 5: We think that the courses will be very difficult for us.

Student 2: Honestly we will not be ready for the first year courses; it will be very difficult for us to study academic courses in English.

7.2.2.3. Comparing GES Tests to Other Tests

In one group, few students maintained that the GES test results could not be compared to IELTS as the GES tests evaluated what had been studied while IELTS evaluated English language proficiency as the following extracts show.

Group 12

Student 1: Tests are necessary but let us not forget that they are not accurate most of the times. I know English language specialization graduates who did the IELTS and scored well because they were trained well or scored badly because they were not trained well. In the IELTS, the results are not always accurate.

Student 5: We cannot compare the IELTS with our exams; our exams are based on what students have learned while the IELTS is based on general language skills, so you cannot compare them.

In three other groups, comparisons of the teaching style, assessment instruments and students' language level improvements were made between the FP in CAS and FP in other institutions. Many students seemed to believe that the teaching style and students' language skills progress were better in other higher education institutions

than they were in CAS. The students seemed to assume the reasons behind better FP in the other institutions were the stricter teachers, richer curricula and better defined assessment structures.

Group (2)

Student 8: If you compare the students who study the foundation programme in the Higher Colleges of Technology and the students who study it in our colleges [CAS] you will find that the English language improvement of the students in the Colleges of Technology is quicker and better.

Group (4)

Student 1: In other higher education institutions, students are handed their assessment schedules by the beginning of the semester, we do not know when the exam will be until a week before.

Student 4: All course plans and changes should be delivered to the students, we should know.

Student 5: I agree.

Student 6: It was us who told our teacher that we had a midterm exam.

7.2.2.4. GES tests Consistency in Measuring Students' Performances

As has been indicated earlier, the students tended to consider the tests and CA as two distinct types of assessment that evaluated different language aspects using varying marking systems. The tests were believed to be consistent measures of language performance generally, but their limited time can turn them into unreliable measures. Many students seemed to think that tests assessed their language skills at points in time which resulted in their ineffectiveness in revealing a complete picture of their actual language proficiency.

7.2.2.5. The Consequences of GES tests

When asked about the fairness of the assessment instruments and importance of passing the FP assessment, the students' responses varied from arguing that the assessment was very fair and passing was very important to claiming that assessment was unjust and passing FP assessment was unimportant. In almost all of the focus groups, FP assessment seemed to be regarded as unfair because of the distribution of scores on test tasks, type of test tasks, or inappropriate curriculum. It seems that the meaning of the concept "fairness" did not only include tests' qualities but was stretched to include course curricula.

Group (3)

Student 12: Tests are not fair; they test grammar more than the other skills. Most of the scores are on grammar and since we are weak in grammar we loose a lot of scores in the tests.

Student 13: I felt that the test let me down. I was depressed because of my low scores.

Group (8)

Student 4: The distribution of the scores is not fair at all. The scores are divided on two instruments. We need more chances and more activities to show our language abilities.

Group (12)

Student 8: We are all agreeing on the principle that the assessment is not fair for this curriculum. Until the curriculum is right and adjusted then the assessment might be fair.

The significance of passing the FP assessment was considered differently in the focus groups. Most of the students believed that they would definitely pass the FP and refused to consider the possibility of failing. For few students, considering the consequences of failing triggered negative social and psychological connotations such as: shame and depression.

Group (3)

Student2: I will feel depressed.

Student1: I will drop out of the college.

Group (4)

Student1: A shame

Student2: It is a shame because we will carry a stigma that we failed in the first and easiest year in the higher education system.

Group (6)

Student5: We will not fail. We started right and we will end up right.

Student6: Not all people can afford studying out of the college, so it is very important to pass.

From the extracts above, it can be inferred that the students' feeling of tests unfairness might have resulted from the structure and content of the tests themselves more than from the tests' associated negative social or educational consequences

7.2.3. Students' Perceptions of AES Continuous Assessment

Like the previous section, this section presents the results of the student focus groups on AES assessment categorised by the aspects of validity namely face validity, content validity construct validity, reliability and impact. Categorizing the evidence from the focus groups is intended to facilitate understanding the students' views on FP assessment effectiveness. These categories were not used to imply a view of distinct types of validities rather they were used to present validity evidence on a clearer way.

7.2.3.1. What AES Continuous Assessment Measures

Though the report and presentation were considered as good assessment instruments *per se*, many students seemed to believe that CA was not suitable for everybody and it did not fully reflect their language skills. Two main reasons for this belief recurred in the focus groups. The first reason was that writing and presenting could result in a performance inhibition caused by students' personal traits (e.g., low confidence and shyness) or learning styles (e.g., auditory, visual, and kinaesthetic). The second reason was the lack of proper guidance, training and practice on writing and presenting (e.g., different criteria sets used by different teachers). The intertwining of the students' opposite feelings of appreciation of CA's role in FP assessment and worry of its shortcomings are apparent in the following discussion.

Group (1)

Student 3: But if you were a silent student by nature, presentations and other oral means might not be just in terms of assessing students levels.

Student 8: I think it should be looked at as a comprehensive thing, I mean assessment. The tests with the presentations complement each other in terms of assessing students English language levels and showing their abilities. Some students are more capable of undertaking the exams while the others are more capable of presenting so the various ways of assessing the students give a fair chance to all.

Student 5: Still presentations influences confidence and does not show the real level.

7.2.3.2. The Content of AES Continuous Assessment

Few students in three groups doubted the content of the AES assessment (i.e., essay and presentation) and said that the essay was too complicated for their English language levels and research skills; and that they sometimes intentionally plagiarised or cheated in other ways. These discussions went as follows:

Group (7)

Student 12: They [the teachers] teach us steps on how to write an essay but never ask us to practice. We need to practice in class or out but the teacher always says that he will not mark our work. How do they know that the end result is my own work if they do not see samples throughout the semester? Last semester I asked my sister to write the essay for me and I will do the same this semester because I simply cannot write it though I know the steps.

Student 1: A student here said that she asks her sister to write her essay, other students download bits and pieces from the internet. I personally did it last semester and the teacher did not know it. And the teacher commented that my essay was a good one, why did not she discover that it was downloaded from the Internet?

Group (11)

Student 2: But there is a huge chance for cheating in the writing project too. They can get former students projects and present them as their own or download things from the Internet.

Student 6: This is the students' problem if they want to harm themselves by plagiarizing, they can do it. And if they want to learn they can do it too. But the fact is that you use so many skills in the essay and the presentation including, listening, speaking, reading and writing. Let us not forget that there are other skills like interviewing, summarizing, planning or looking up information in the internet.

7.2.3.3. Consistency in Implementing AES Marking Scales

As has been pointed out earlier, most of the students seemed to be aware of the fact that the marking scales will be used AES assessment but not of how they will be used. It was apparent from most focus groups that many students felt that the marking scales were inconsistent in how they were implemented or interpreted.

Group (7)

Student 12: But teachers differ in terms of the criteria they use to assess the students. We know that other groups are told different things about how they will be assessed in the essay. This is wrong, we are not assessed equally. All students should be given the criteria at the beginning of the semester before starting to write the essay or preparing for the presentation.

Group (6)

Student 8: It [marking the essay] depends on the way of a student writing, if you write well you will get good marks. The teachers highlight students mistakes in drafts and then the students get better scores after that.

Student 5: It depends on the words number, grammar, vocabulary and ideas.

Student 3: We have not been told about all of this.

Student 4: Different teachers explain to their students how they will be assessed differently. For example we do not know about how the scores are distributed in the essay.

7.2.3.4. The Feedback Given in Continuous Assessment

A recurring theme in focus groups was the lack of teacher feedback offered in the essay and presentation. Generally in most of the groups the students expressed dissatisfaction with the amount and nature of feedback provided; and argued that the appropriate feedback could improve their language skills. It was claimed that sometimes the feedback imparted was ambiguous, negative, delayed or non-existent. The subsequent extracts display some of the students' comments on teacher feedback.

Group (8)

Student 2: there is no feedback at all, we do not see our scores and we do not know how we are doing so far, we just wait until the end of the semester and wait to see the result at the end. We should have been given some feedback to lead us on what we should be learning and how we should do things right.

Student 6: we need more quizzes more things to tell us about our levels and guide us in learning. But most importantly we need feedback on the tools.

Group (7)

Student12: Teachers when marking, they only underline they do not write what is wrong and it becomes a guessing game for us to know what is wrong. Most of the time we give up and we do not know what is wrong.

Student 2: We do not care about attending the classes, we do not learn anything, and the teacher does not care about teaching.

Nonetheless, only two groups of the fifteen groups seemed to be satisfied about the feedback received on their essays and presentations and tended to attribute their satisfaction of the feedback offered to their teachers' teaching styles.

Group (1)

Student 5: It depends on the teachers, if the teachers help the students in giving them good feedback, for example when a teacher allows students to write two drafts before submitting a paper and gives them good feedback, then a student will do well. Teachers here are very cooperative in telling the students what and how they should improve their essays.

Group (2)

Student 9: Last semester was better in many levels, we did plenty of assessment activities and there was enough feedback on our levels. The teachers were better and more involved in teaching us everything.

Student 1: This semester the way we are going to be assessed is not clear. No one explained to us. The course plan seems to be not stable and the teachers seem to be not sure of how and what the assessment will look like.

7.2.3.5. The Consequences of Continuous Assessment

The essay and presentation were described as subjective and unfair by most of the groups. Teaching styles, marking scales, scores distribution and availability of resources were all considered factors that participated in characterising CA as being unfair.

Group (8)

Student3: The fairness of assessment depends on the teachers. Some teachers are unfair in marking the exams. For example sometimes we cannot revise the exam results with the teacher or discuss any concerns some of them get angry when you try to discuss the scores with them.

Group (2)

Student 5: I cannot understand how it is 50% of the total score in the Academic English Course is on a 5-minutes presentation. It is not fair.

Student 2: I agree, it is not fair.

Group (3)

Student 8: We are assessed according to the resources we use; however, the college library is very poor in terms of books or other resources that could be used in our research.

Student 11: In the General English course the speaking component is worth only 12% of the test marks. This is a very low percentage given that there are some students who are really good in speaking. The speaking component should be given more weight. This is not fair.

Only one group considered AES assessment to be fair not because of its qualities but because of the teachers who were regarded as being fair.

Group (1)

Student2: With our current teachers, I think it [AES] is fair and they [the teachers] are fair.

Student5: It depends on the teachers, if teachers help students in giving them good feedback, for example when a teacher allows students to write two drafts before submitting a paper and gives them good feedback, then a student will do well.

7.3. Results of the Teacher Interviews

Once the teachers filled in the questionnaires, they were asked to participate in a 30 minute semi-structured interview. Of the 27 teachers who participated in filling out the questionnaires, only 19 teachers agreed to be interviewed. Ten were from Sur College and nine were from Rustaq College. Eleven of the participants were male teachers and eight were female teachers. Table 7.2 displays the college, gender, nationality, taught course (i.e., GES, AES) and qualifications of the participants.

Table 7.2. College, Gender, Nationality, Taught Courses and Qualifications of Teachers in Phase1 Interviews

Teacher	**College**	**Gender**	**Nationality**	**Taught Courses**	**Qualification**
Teacher 1	Sur	M	British	GES/AES	BSc
Teacher 2	Sur	M	Canadian	GES	BA
Teacher 3	Sur	M	British	GES	BA
Teacher 4	Sur	M	Syrian	GES	PhD
Teacher 5	Sur	M	Greek	GES	BA
Teacher 6	Sur	F	Omani	AES	MA
Teacher 7	Sur	F	American	AES	BA
Teacher 8	Sur	M	Omani	GES	MA
Teacher 9	Sur	M	Omani	GES	MA
Teacher 10	Sur	M	British	GES	BSc
Teacher 11	Rustaq	F	British	AES/GES	BA
Teacher 12	Rustaq	F	British	AES/GES	BA
Teacher 13	Rustaq	F	Romanian	AES	PhD
Teacher 14	Rustaq	F	Omani	GES	MA
Teacher 15	Rustaq	F	Omani	GES	MA
Teacher 16	Rustaq	M	Iraq	GES/AES	BA
Teacher 17	Rustaq	F	British	AES	BA
Teacher 18	Rustaq	M	British	GES	BA
Teacher 19	Rustaq	M	British	GES	BA

Following the coding and analysing procedures discussed in Chapter 4, four main themes emerged from the interviews: (1) uncertainty about assessment instruments weightings and scales, (2) perceptions of the effectiveness and impact of GES tests, (3) perceptions of the effectiveness and impact of AES continuous assessment, (4) and perceptions of FP students. The second and third themes were divided into subthemes. These themes are intended to respond to the study questions posted at the beginning of this chapter.

7.3.1. Uncertainty about the Assessment Instruments

When the interviews were conducted a month before the end of the semester, most of the interviewees were aware of the types of assessment instruments applied in that semester in both the GES tests and AES assessment. However, four of the 19 interviewed teachers seemed unaware of the scores distribution on the two GES tests which was 40/60. Occasionally, different marks' allocations were mentioned when discussing the effectiveness of FP assessment, as appears in the three extracts below.

> Teachers 1: Well we have a midterm exam which is 30% of their final grade and there is going to be a final exam which is going to be 70% of their grade which is heavily weighted.
>
> Teacher 7: I am doing the Headway book meaning all of their scores come from a midterm and a final: 50% midterm and 50% a final test. The other class is a writing class so 50% of the scores is for a writing project and 50% of the scores is for a presentation.
>
> Teacher 14: I would say 50% for the final exam which is very important and 50% for the midterm.

Teachers' uncertainty was not limited to the weightings of GES tests but included some aspects of the AES assessment. One teacher stated that a 300-words essay was too long for her students to handle, whereas, actually, the essay instructions mentioned that the students should write a 500-word essay. Several other teachers seemed uncertain about the criteria of the marking scales and how they should be implemented.

7.3.2. Teachers' Perceptions of GES Tests

7.3.2.1. What GES Tests Assess

The interviewed teachers seemed to have various views about the GES tests. Most of the interviewed teachers indicated that the tests were generally suitable for assessing students' English language proficiency and evaluating their progress. However, about a third of the teachers showed concerns with regard to the discrepancy in the difficulty levels between the test tasks and taught materials. Similar to the students' view, they seemed to think that the tests' difficulty levels were not appropriate for

the student language levels in that the test tasks were believed to be challenging for the students. These concerns are manifested in the extracts below.

> Teacher1: The test is correlated with the materials covered in the books. The problem is that there are deeper issues here; principally the materials and syllabus do not match the students' levels very well.
>
> Teacher7: In theory the test levels should match the students' levels, but because of the reality of the level of the students that is not really what is going on the classroom because it is not possible. So if you look at the exam paper and curriculum they match but what happens is that they do not always match the students' levels.
>
> Teacher12: You know it [the test level]. It is like getting a manual about fixing a computer and applying it to making bread.

Despite these concerns, the GES tests were believed to be a more accurate measure of students' language proficiency than was the AES assessment for the former's perceived objectivity. One teacher stressed that AES "assessment bends towards more students passing because of its sympathetic and subjective marking whereas the final exam is less so, and fewer students get through". In the extract below, another teacher referred to AES assessment as a piece of cake that everybody could easily accomplish, unlike the GES tests.

> Teacher 16: There is more I agree with using the midterm and final tests because at least there is a fixed thing you can follow. Unlike AES assessment, no body fails in this assessment, all of them get the highest marks. It is a piece of cake for them.

7.3.2.2. Perceived Need for More Quizzes

The majority of teachers commented that there was a need for more short quizzes or tasks distributed throughout the semester "to know where the students are and what they need". They seemed to endorse using multiple short quizzes during a semester as they "show what actually they [students] can do", "structure the course", "build a good relationship with your students", and "end up with better students' attendance".

Furthermore, most of the interviewed teachers compared the GES assessment in this semester which evaluated students' language proficiency using a midterm test and a

final test to last semester which assessed the students' language skills using weekly quizzes. They maintained that last semester had been better in terms of the feedback the students and teachers received from the quizzes. The quizzes were also believed to contribute to improving students' attendance rate. Furthermore, many of the teachers argued that administering two tests (i.e., a midterm test and a final test) were not enough as the only assessment instruments in a ten hours course. They claimed that two tests did not provide enough or suitable feedback to the students and teachers most of whom had devised their own classroom tasks to evaluate the students' progress. These views are depicted below.

> Teacher 7: Having more quizzes maybe not as many as we had last semester but having more opportunity for them throughout the semester instead of having everything at the end.

> Teacher 9: Quizzes are good, time consuming and paper consuming but they are good. Both sides are benefiting from it, the students and the teachers even the midterm is good.

Though the teachers seemed to support increasing the number of the assessment instruments many were apprehensive of using quizzes similar to the ones that had been used in the previous semester. They seemed to indicate that the quizzes used last semester were not thought through or planned properly and future use of the quizzes should be premeditated and be more reliable.

> Teacher 11: The tests [referring to the quizzes used] we were sent last semester, for the reading skill, were exactly the same passages from the book and they have not been bothered to devise new questions, so they were exactly the same questions that the students had in the book.

> Teacher 12: Last year, it [quizzes] all came from Muscat, they [students] set down a weekly quiz. It was so stupid that it was embarrassing.

> Teacher 2: Well I think here should be an assessment team that puts together, that creates a good level of assessment as the assessment at the end of the book are good but they are only relative to the book they are not relative to the class itself … There needs to be some revision done.

7.3.2.3. Unavailability of Past Exam Papers

Many teachers expressed that it was essential for the students to see examples of previous tests to prepare them for the midterm and final tests, yet, no samples or mock tests were provided. This was justified by a possible future recycle of the test tasks. The teachers' views on this matter varied and reflected frustration, understanding or support of the resolution to withhold previous tests from teachers and students. Regardless of this resolution, it was widely felt that sharing the previous tests with the students was a necessity to assist them to avoid confusion or distraction by the test format.

> Teacher 9: The outline of the frame of the exam, not necessarily the questions, only the frame. It is something similar to the IELTS, if you know what you will be faced with, and then you will be prepared in a better way for it.

> Teacher 17: There [in the tests] tend to be more grammar but that where the students feel less confident…Although they are not on the final exact format, it would be good if we could use previous tests with the students.

> Teacher 2: We all struggle to prepare the students for what they are going to do. We do not have access to the previous exams, well I have as I am a coordinator but the teachers do not.

7.3.2.4. Impact of GES Test: Passing to the First Year (FY)

The majority of the interviewed teachers seemed to believe that some students were passed up to FY regardless of the predicted inability to function in FY and fulfil its language related requirements. It was implied that the colleges tended to lower the bar so more students can start FY courses. It was also suggested that overlooking plagiarism in the AES assessment could have contributed to allowing incapable students to join FY.

> Teacher 1: Certainly there is a large number of students who pass every year who should not pass. Officially they should pass the foundation with an IELTS score of 5 -I do not know where this come from- but they are nowhere near 5 some of them will take only 2.5… that threshold is lowered to say 30% to allow more students to escape.

> Teacher 7: The majority of them will pass but I do not think will realistically have the skills for the First Year. They will pass but they will be low students in

First Year and will be more difficult to work with. I know how the college works so I am sure that they will be moved up.

Teacher 10: I understand it as being basically political I suppose and how the college want to be seen. Teachers are not involved in the up lifting in any shape or form but the scores have been adjusted to fill the classes.

Teacher (12): Everyone passes.
Interviewer: How do they pass?
Teacher: They do not look at plagiarism; this is the reality of it.

7.3.3. Teachers' Perceptions of AES Assessment

There was a general satisfaction with the AES assessment instruments namely the essay and presentation. The majority of the teachers liked the fact that these instruments (1) focused on language skills such as writing and reading and non-language skills such as organising and researching, (2) and corresponded with the objectives of the AES course. A teacher stated that "I like teaching them the presentation and reports because I can help them in structuring sentences and paragraphs". Another teacher said that students responded to the feedback provided in the essay and presentation well and this contributed positively in their overall learning process. A third teacher highlighted the stress free environment the essay and presentation provided the students with. Such an endorsement to using essays and presentations as assessment instruments was also implied in the subsequent two extracts.

Teacher 13: [In reference to essay writing] it is a good assessment tool if it is their own work and in many cases students are making an effort.

Teacher 17: That [assessment] happens through the means of a presentation and a report which is the same type of assessment that goes with them into the First Year and Second Year. So it seems to prepare them for that type of assessment and the students seem to enjoy it. It is something that they can work towards and do independently.

7.3.3.1. Concerns about AES Continuous Assessment

Though most of the teachers seemed to emphasise the vital role played by CA in evaluating and developing language skills, they raised concerns about the (1) high difficulty level of the CA tasks especially the writing one, (2) students' tendency to plagiarise, (3) and inconsistency in using the marking scales in and across the

colleges. Firstly, asking students to write a 500-word academic essay seemed to be considered challenging for most of the students who could not actually write proper paragraphs according to their teachers. Two teachers expressed this concern as follows:

> Teacher 6: The writing project has some problems because the students are still learning at the sentence level … Through following the students in the classroom; I found that the students are not ready for the 300 words essay.

> Teacher 12: You are asking students to write a 500-words project and most of them do not know what a paragraph is … but to force somebody and to tell them that they have 50% of their mark on 500-words essay and most of them cannot even write a coherent sentence, how demoralising is it to be asked something that is impossible to do.

Secondly, as a result of the perceived challenging writing task, several teachers seemed to believe that students tended to plagiarise. A teacher pointed out that "the project allows you to assess their writing skills and research skills but there is a big issue of plagiarism". The issue of students' plagiarism was frequently referred to in teachers' discussions of the AES assessment. The following extracts exemplify the contexts in which teachers talked about it.

> Teacher 13: In certain cases some students write something in Arabic and put it through Google translator. And I cannot find these texts in Google because they do not exist or they take an Arabic text from a web site and again they put it through Google translator and in both cases you cannot tell if the work is plagiarised.

Thirdly, the reliability of marking the essay and the presentation was questioned by a small number of teachers who argued that there was a lack of consistency in using the marking scales and doubted the reliability of the markers who sometimes, as was believed, had not used the marking scales at all.

> Teacher 17: [talking about the presentation scales] some teachers might use them and some teachers might not, I think most teachers use them.

7.3.3.2. Perceived Need for More Continuous Assessment Tasks

When interviewed about the effectiveness of CA, many teachers seemed to believe that more assessment tasks were needed. It was explained that sometimes students studied only what was covered by the assessment tasks; and that they had different learning styles so different assessment tasks were hoped to capture these differences. In the extract below, two teachers provided two rationalizations for using different types of assessment instruments.

> Teacher 7: I think that it would be better if the students had the more opportunity to show what actually they can do and more different ways to show the skills and all of the students have different learning styles and so some of them they cannot test well but can do amazing presentations.

> Teacher 13: Well you see, in Foundation, the only tool I am using is the essay and the presentation and they can get to that quite easily.

7.3.3.3. Comparing CA to Tests

Similar to the students' attitude, most of the teachers seemed to prefer CA as an assessment instrument when compared to tests. The justification for this preference ranged from the extended time scale and the stress free environment CA offered to the opportunity to develop and learn language skills while being evaluated. Some of these views are expressed in the following extracts.

> Teacher 2: The assessment is more forgiving in the sense that you have a one bad day you can make it up with another exam. Whereas the midterm, if you are already a good student and you know that you could pass the exam, your participation in class might not be as much as if you were tested every few weeks.

> Teacher 7: They get very nervous when they get to the exam time and from what I have seen here from my past students they test lower than they work in the classroom.

> Teacher 10: I think it is CA which is more reliable measures of students' abilities than isolated prompting in time tests.

7.4. Discussion

This section brings together the results of the student focus groups and teacher interviews to reach a better understanding of the perceived FP effectiveness following Hamp-Lyons's suggestion of involving the views of both the teachers and students in the process of evaluating assessment practices and their impact.

> It is not enough to evaluate tests from our own perspectives; neither is it enough to evaluate them by including teachers' perspectives… Many more studies are needed of students' views and their accounts of the effects on their lives of test preparation, test taking and the scores they have received on tests (1997, p. 299).

This section discusses the results in five sub-sections that summarise the main findings and attempt to answer the study questions listed at the beginning of this chapter. It starts with investigating the uncertainties about FP assessment reported by both the students and their teachers; then it reports on the perceived effectiveness of FP which is immediately followed by discussing the expressed need for more assessment tasks. The fourth and fifth subsections discuss the feedback and social consequences of FP assessment as part of exploring FP assessment impact.

7.4.1 Uncertainties about the FP Assessment Elements

In the interviews and focus groups, several teachers and most students seemed uncertain about specific aspects of the AES and GES assessment instruments. Some of these aspects were shared by both the teachers and students such as: the weightings of the assessment instruments and the criteria of the marking scales. Other aspects were either teacher specific (i.e., essay length) or student specific (i.e., test sections). Empirical evidence have suggested that the students' understanding of assessment requirements might well be different to that of their teachers' as Green (2007) indicated in reference to Weir and Green's study (2002). In line with this suggestion, this study found that indeed students expressed a less certain understanding of what was required by the assessment activities than that of their teachers'. In AES assessment, for instance, most students complained about the lack of information on some aspects of which their teachers seemed very well aware.

Students' and teachers' uncertainties about aspects of FP assessment could be also referred to the unavailability of sample or mock tests. Most of the teachers reported that past exams were not accessible for them or their students and consequently they

were not completely aware of the exams' structure and contents. Rea-Dickins (1997) asserted that in centralised systems where teachers were not involved in assessment development, they could be not "prepared sufficiently for the task of implementations" (p.308). In the context of the current study, though GES tests were written by individual assessment coordinators from different colleges, the tests were not distributed to the rest of the teachers several of whom were novice in the Colleges and had never seen these tests before. Understandably, several teachers and many students seemed ill-informed about FP assessment.

Furthermore, when considering the larger picture, three focal issues could be contributing to the reported lack of information about some aspects of FP assessment. Firstly, there was a change in assessment instruments between the autumn and spring semesters. Phase 1 of this study was conducted in the spring semester of the academic year 2010/2011. The assessment that was used in the GES course had been changed from being a series of quizzes and a final test to a midterm test and a final test. This change entailed a modification in the weightings of GES assessment instruments of which many students and few teachers were not aware. Assessment in the AES course remained as it was in the autumn semester; therefore, the students' and teachers' informed knowledge of AES assessment weightings was rather expected. Nonetheless, the criteria of the marking scales used in the AES assessment were constantly debated amongst the students. This may suggest that the marking scales were not shared or explained to the students appropriately. In a review of assessment studies in the field of education, Gipps (1999) argued that there should be more opportunities to discuss the assessing criteria with the students to help students make sense of assessment requirements.

Secondly, the contradictory information about the FP assessment that appeared in the online documents in coordinators' website could have contributed to this confusion and lack of clarity shared by the teachers and students. In this study, most of the assessment related documents provided for the FP coordinators and assessment coordinators were analysed revealing that some documents were not updated with the latest changes about the assessment instruments' weightings or scales. This mismatch

amongst sources of information in different documents might have led the teachers and their students astray.

Thirdly, the reported high teacher turnover rate in both colleges could have contributed to the teachers' expressed uncertainty about aspects of FP assessment. One of the FP coordinators expressed the difficulty of keeping all teachers on-board about the recurrent changes in FP assessment because of the high teacher turnover rate.

7.4.2. FP Assessment Effectiveness in Student and Teacher Perceptions

Both most students and teachers seemed less satisfied with the GES assessment (i.e., tests) than they were with the AES (i.e., presentation and report). The content of the GES tests was severely criticised by both the teachers and students. The teachers focused on the mismatch in levels between the taught materials and the tests used. The students emphasised the difficulty that they faced in the grammar, reading and listening sections of the mid-term test. They elaborated that the reading topic was new; the grammatical rules were not all covered in the course; and the listening genre had not been introduced to them before. Messick (1996) suggested that "to facilitate positive wash back, the assessment must strive to minimise construct underrepresentation and construct-irrelevant difficulty in the interpreted scores" (p.245). In this study, both teachers' and students' perceptions of the difficulty level and content of the tests indicate that the tests have reflected aspects of the "construct irrelevant difficulty". This issue will be further discussed in Chapter 11.

Though AES assessment was generally positively viewed by the students and their teachers, they both made comments signalling its problematic content and construct. They raised three similar concerns about the essay: (1) high difficulty level, (2) plagiarism, and (3) variability in implementing marking scales. In focus groups, some students admitted to committing plagiarism because they found the essay very difficult for them to write using their own words. Students' interactions with the assessment tasks have been identified as a parameter in understanding students' performances and difficulty of assessment tasks (Bachman, 2002). Several teachers reported incidents of plagiarism and attributed them to the difficulty level of the essay task. In her review of studies on educational assessment, Hamilton (2003)

discussed a number of studies that investigated cheating in tests; one study (Jacob and Levitt, 2003) found that the cheating instances increased when the tests were high-stakes. Another study on students' perceptions of plagiarism in higher education found that students sometimes perceived plagiarism as "a strategy for coping with the demands of higher education level work and the pressure to succeed" (Ashworth, Bannister, & Throne, 1997, p. 194). A similar perception was documented in other studies in the field of second language learning and assessment (Currie, 1998; Pecorari, 2003). The findings of this study conform to the findings of the studies that have recognised task difficulty as a contributing factor that influences students performances (Bachman, 2002); and considered it a principal factor in resorting to plagiarism (Hamilton, 2003).

Furthermore, the difficulty level of the essay task was not the only element of AES assessment criticised, the students and their teachers expressed their apprehension of the inconsistency in implementing the marking criteria. This concern seems to match similar concerns documented in several studies on performance assessment (Brindley, 1998, 2001; Hay & Macdonald, 2008). Brindley (1998) reviewed a number of articles and identified numerous problematic issues with the validity of the scales used to mark students' performances; he categorised them into political, technical and practical. He asserted that "subjective judgements of language performance are likely to show a good deal of variability" (p.65). Addressing this concern, Gipps (1999) advised that rater inconsistencies should be minimised to reach a better reliability especially in high-stakes assessment tasks. Given the high-stakes nature of FP English language assessment, and the concerns raised by both the teachers and students about inconsistency in implementing marking measures, there seems to be an urgent need for implementing the standardization and moderation procedures discussed (see Section 5.3.3).

7.4.3. Perceived Need for More Assessment Instruments

Regardless of the previously mentioned concerns about the effectiveness of FP assessment, both the teachers and students tended to express the need for more assessment instruments. The students' and teachers' declared a request for more assessment tasks can indicate and be linked to the need for more feedback.

Administering additional assessment tasks and feedback might appear unrelated but they actually are when considering the findings of the studies conducted on feedback suggesting that summative assessment provides less feedback than does formative assessment. Brindley (1998), in a comparison of summative and formative assessment, stated that the former was more suitable for the purposes of policy makers and educational bureaucrats for its skimmed aggregated details, while formative assessment provided detailed and elaborated feedback. Broadfoot (2007) identified the purpose of summative assessment as to "sum up to progress of an individual in relation to some given criterion" (p.110), and the purpose of formative assessment as to provide "information to be used as feedback to modify the teaching and learning activities they are engaged in" (p.111). In the higher education context, there seems to be a move towards less formative assessment and more summative assessment with late or insufficient feedback (York, 2003). Revisiting the findings of this study, it could be noticed that both GES and AES assessment instruments might be considered as summative with regard to the time and type of feedback with which the students are provided. Though the students received some sort of feedback on the first and second drafts of the essay; this feedback, as considered by the students, was occasionally detrimental, late or insufficient. Even, the few students who seemed to believe that feedback on the essay was appropriate attributed their satisfaction of the feedback provided to having a good teacher.

Hamilton (2003) reviewed a number of studies that provided evidence of better students' performance when more feedback on how to improve performance was given to them. Likewise, reviews on the effectiveness of feedback showed that it varied based on different aspects (Bangert-Drowns, Kulik, & Kulik, 1991). Recent reviews that focused on what the feedback was about (i.e., task, processing, regulatory) found that feedback was most effective when it attended self-regulation (Hattie & Timperley, 2007). A conforming finding was reported by Black (2003, as cited in Broadfoot, 2007), who asserted that 'task-oriented' feedback enhanced the 'learning power' of the students and enabled them to take control of, and encouraged them to get involved in their own learning.

7.4.4. Comparing CA to Tests

Similar to the results obtained from the questionnaires, both the students' and teachers' views seemed to generally prefer CA more than the tests for several reasons. The teachers attributed this preference to CA characteristics (i.e., less stressful, more reliable than tests and students better chances to perform well in CA). The students' preference of CA seemed to be propelled by their appreciation of the process of learning that takes place in the evaluation process of the students' language proficiency. This apparent students' appreciation of learning through assessment is in line with and reinforces the voices calling for "assessment for learning" as a way forward in assessment for its ability to improve students' performances as supported by empirical evidence (Broadfoot, 2007).

7.4.5. FP Assessment Impact: Passing to the First Year

The results showed that the students seemed to be very confident of passing the FP assessment but revealed their concern that their language level would be lower than what is required for FY Study. Similarly, most of the teachers tended to believe that many students were allowed to pass FP regardless of their unsuitable level of English language for the FY courses. The FP results in 2011 showed that more than 90% of the students in both colleges passed to FY, though, the teachers generally expected less than 80% of their students to pass. In Sur College 92% of the FP students successfully passed and in Rustaq College 97% of the FP students passed. The students' and teachers' view of FP assessment inability to fulfil the role of filtering the linguistically able students to study in FY, could pose a threat to the validity of FP assessment. Messick (1996, p.245) asserted that "validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. Hence what is to be validated is not the test or observation device *per se* but rather the inferences derived from test scores or other indicators". From the teachers and students' understanding of the meaning of FP assessment scores, and from Messick account of assessment validity, it could be argued that FP assessment shows signs of problematic validity; more evidence supporting this argument will be presented in Chapter11.

7.4.6. FP Assessment Impact: The social aspect

An unexpected result obtained from the focus groups and interviews was the relatively moderate to non-existent social impact of FP assessment considering its high-stakes nature. Failing in FP could mean that students lose their scholarships to study at CAS or become suspended for one academic year during which an accredited proof of a specific language level should be attained from a recognised private language teaching institution. However, not only very few students expressed that failing in FP assessment could entail a negative social stigma, most of them seemed to be confident that they would pass and did not show any concern of failing in FP. Shohamy (2001) explained that there are a number of factors that could contribute to understanding the consequences of a test like language status, purpose of assessment, format of assessment and low/high-stake nature of tests. Though all of these factors when considered in the context of FP assessment predict a strong negative social impact, the findings of this study arrived at a different conclusion. A possible logical explanation for this finding is what the teachers indicated about inconsistent implementation of FP assessment marking criteria.

7.5. Summary and Concluding Rescores

This Chapter presented the findings obtained from the student focus groups and teacher interviews in Phase 1. Using thematic analysis, five common themes emerged from the data set: uncertainties about FP assessment, effectiveness of FP assessment, perceived need for more assessment tasks, comparing CA to tests, and impact of the FP assessment.

The findings on the first theme revealed that the students and teachers were uncertain about several essential aspects of the FP assessment. It is important that, in this kind of high-stakes centralised assessment, that the structure and nature of assessment instruments are made clear to both students and teachers to minimise any uncertainties since the majority of the teachers were not involved in producing CA or tests. The lack of clarity of the assessment criteria and structure continues to occur in students and teachers comments in the second phase of the study in Chapter 9.

The students' and teachers' perceptions of the effectiveness of FP assessment attacked several aspects of FP assessment such as: difficulty level, content and inconsistency in marking. Similar and sometimes more elaborate findings are reached in the other chapters which will be all joined to construct a comprehensive argument about FP assessment validity in Chapter 11.

The findings on the FP assessment feedback revealed insufficiency of the amount and type of feedback offered to students due to the summative nature of the assessment instruments used. This should be considered and FP assessment instruments should be redesigned to maximise formative feedback while keeping the summative nature of the assessment for accountability purposes. The redesign could entail increasing the number of assessment instruments or merely adjusting the current ones to provide additional feedback.

Students' and teachers' perceptions of FP impact can be generally encapsulated by the educational impact of moving up linguistically unready students to academic study and the almost non-existent social impact. Both the students and teachers affirmed that many of the students who pass FP assessment were unready to undertake academic courses in English. They also seemed to believe that FP assessment imposed slight if non-existent negative social impact. This finding conforms to similar ones obtained from the questionnaires in Chapter 6 (see sections 6.2.3 and 6.3.3). Similarly, students mentioned several difficulties they faced in the FY study and their teachers argued that many of the students were unready to undertake academic courses (see Sections 9.2.2 and 9.3.2).

Chapter 8: The Results of the Student and Teacher Questionnaires in Phase 2

8.1. Introduction

This chapter reports on the results obtained from the student and teacher questionnaires carried out in the second phase of this study (i.e., autumn 2011). The purpose of using these questionnaires was to attempt to answer the study questions in Box 8.1. The first section of the chapter presents the results from the student questionnaire displaying the students' responses to the individual items. This is followed by a presentation of the average responses to each topic of the student questionnaire. In the same section, the significant differences in responding to the questionnaire topics across the groups are identified. In the second section of the chapter, the results obtained from the teacher questionnaire are presented in a similar order to that followed in displaying the results from the student questionnaire. In the third section of the chapter, the results of both questionnaires are compared and discussed in the light of related previous studies and the findings obtained from other methods presented in the current study.

Box 8.1. Study Questions Addressed by the Results Obtained from the Student and Teacher Questionnaires in Phase2

4. How did the stakeholders understand the relationship between the student performances in the English language assessment and their performances in the academic courses' assessment?

4.1. What were the teacher and student perceptions of issues related to the design, marking and impact of the English language assessment?

4.2. How did teachers and students think language accuracy should be considered in assessing academic assignments?

4.3. What were the students' and teachers' perceptions about the importance of the predictive validity?

8.2. The Student Questionnaire

8.2.1. Demographic Characteristics of the Participants

The students (*N*=184) who had participated in the first phase of the study were contacted to participate in the second phase; A total of 176 students agreed to respond to Phase 2 student questionnaire; the other eight missing participants either did not pass the Foundation Programme (FP) assessment or expressed their reluctance to be involved in the second phase. In this phase, students were taking First Year (FY) academic courses. [To remind the reader, Phase 1 of the study was conducted in the Spring Semester of 2011 which started from February 2011 to June 2011 and targeted FP students; Phase 2 of the study was conducted in the Autumn Semester of 2011 which started from September 2011 to January 2011 and targeted FY students.]

The demographic characteristics of the participants in the second phase were as follows; 122 participants (69.3 %) were from Rustaq College and 45 participants (30.7%) were from Sur College; 116 participants (64.7%) were females and 60 participants (35.4%) were males. The sample distribution by specialization is as shown in the table below.

Table 8.1. The Students Distribution by Specializations in Phase 2

Specialization	***n***	**%**
Information Technology (IT)	48	27.3
Communication Studies (CS)	24	13.6
International Business administration (IBA)	82	46.6
English Language-Education (EL)	22	12.5
Total (***N***)	176	100

8.2.2. Students' Responses to Individual Items in the Questionnaire

A 21 items-Likert scale questionnaire was administered in the second phase of the study to survey students' perceptions of the English language assessment after passing the FP assessment and embarking in academic studies. As discussed in Section 4.4.2.1., the questionnaire items were organised into six topic areas: *Dissatisfaction with Language Assessment, Adequacy of Students' English Language for FY Study, FP Assessment Predictive Validity, FY Construct Validity, Impact of*

FP Assessment, Assessing Language Accuracy and Content. The students were asked to select one of five numbers in each of the items. The numbers were used to express a level of agreement or disagreement (i.e., 1= Strongly Agree, 2=Agree, 3= No Opinion, 4= Disagree, and 5= Strongly Disagree) as shown in Table 8.2 below. The process by which the means were calculated and the process by which the items were recoded is similar to the one explained in Section 6.2.2.

Table 8.2. Frequency, Percentages and Means of Responses to the Student Questionnaire in Phase 2

Topic	Item	SA 1	A 2	NO 3	D 4	SD 5	Mean	Mean (Recoded)
Dissatisfaction with Language Assessment	1.1. Assessment on the FP should have allowed more students to proceed to the FY.	39 (22.2%)	27 (15.3%)	44 (25.0%)	36 (20.5%)	30 (17.0%)	2.95	–
	1.2. Assessment instruments should be changed on the FP to better match my English language needs in academic courses.	76 (43.2%)	57 (32.4%)	35 (19.9%)	5 (2.8%)	3 (1.7%)	1.88	–
	1.3. Assessment instruments should be changed in the FY to better match my English language needs on academic courses.	73 (41.7%)	56 (32.0%)	29 (16.6%)	13 (7.4%)	4 (2.3%)	1.97	–
Adequacy of Students' English Language for FY Study	2.1. My English language level is adequate to understand the academic courses and to meet their assessment requirements.	79 (44.9%)	53 (30.1%)	34 (19.3%)	7 (4.0%)	3 (1.7%)	1.88	–
	2.2. I have difficulty understanding my lecturers in the FY academic courses because my English language level is insufficient. (Recoded)	42 (23.9%)	64 (36.4%)	44 (25.0%)	21 (11.9%)	4 (2.3%)	2.39	3.61
	2.3. I have difficulty in expressing my ideas in writing in the academic course assessments.(Recoded)	50 (28.4%)	70 (39.8%)	20 (11.4%)	28 (15.9%)	8 (4.5%)	2.28	3.71
	2.4. I have a difficulty in understanding the reading passages for the academic courses assessments. (Recoded)	38 (21.6%)	62 (35.2%)	50 (28.4%)	21 (11.9%)	5 (2.8%)	2.43	3.29
Foundation Programme Assessment Predictive Validity	3.4. The better a students' English language ability, the better his/her achievement in academic courses will be.	126 (71.6%)	40 (22.7%)	5 (2.8%)	2 (1.1%)	3 (1.7%)	2.14	–
	3.5. I needed more English language courses if I am to perform well in the First Year.	66 (37.5%)	52 (29.5%)	33 (18.8%)	17 (9.7%)	8 (4.5%)	1.39	–
m e n t C	4.6. Assessment instruments in the FY measured my	17	55	49	32	23	2.94	_

	language skills appropriately.	(9.7%)	(31.3%)	(27.8%)	(18.2%)	(13.1%)		
	4.7. In the English language course, teachers assess both my ideas and my language. 4.8.	18 (10.2%)	27 (15.3%)	47 (26.7%)	61 (34.7%)	23 (13.1%)	3.25	–
Impact of FP Assessment	5.4. The assessment and teaching in English creates more employment opportunities for me.	108 (61.4%)	50 (28.4%)	8 (4.5%)	9 (5.1%)	1 (0.6%)	1.55	–
	5.5. Teaching and assessing in English at university level supports my country's status internationally.	90 (51.1%)	52 (29.5%)	12 (6.8%)	12 (6.8%)	10 (5.7%)	1.86	–
	5.6. FP assessment has more negative social consequences to me than FY assessment.	31 (17.6%)	27 (15.3%)	25 (14.2%)	38 (21.6%)	55 (31.3%)	3.34	–
Assessing Language Accuracy and Content in Academic Courses	6.1. Teachers on academic courses should assess students on their written expressions as well as their ideas.	71 (40.3%)	82 (46.6%)	18 (10.2%)	4 (2.3%)	1 (0.6%)	1.76	–
	6.2. I would like to get feedback on both my ideas and my written expression in academic courses.	73 (41.5%)	82 (46.6%)	12 (6.8%)	7 (4.0%)	2 (1.1%)	1.77	–
	6.3. In academic courses, students should not be marked for their English language skills. (Recode)	37 (21.0%)	84 (47.7%)	33 (18.8%)	19 (10.8%)	3 (1.7%)	2.24	–
	6.4. Academic course teachers assess both my ideas and my language.	46 (26.1%)	61 (34.7%)	38 (21.6%)	25 (14.2%)	6 (3.4%)	2.34	–
	6.5. I think that assessment in the academic courses should not require written assignments in English.	21 (11.9%)	36 (2.5%)	55 (31.3%)	39 (22.3%)	25 (14.2%)	3.06	2.99
Assessing Language Accuracy and Content in English	7.1. In the English language course, teachers assess both my ideas and my language.	91 (51.7%)	68 (38.6%)	11 (6.3%)	4 (2.3%)	2 (1.1%)	1.63	–
	7.2. I would like to get feedback on both my ideas and written expression in English language courses.	21 (11.9%)	61 (34.7%)	55 (31.3%)	27 (15.3%)	12 (6.8%)	2.7	–

Two main points can be noticed from the students' responses to the individual questionnaire items. First, though most of the students seemed generally dissatisfied with the FP assessment, their opinions about whether more students should have been allowed to pass seemed to be mixed. In the *Dissatisfaction with FP Assessment* topic, more than 70% of the students seemed to believe that the English language assessment should be changed in both the FP and FY. However, the percentage of the students who expressed agreement with allowing more students to pass FP by lowering exit criteria was actually identical to the percentage of those who disagreed with the same issue (i.e., 37.5%); while 25% of the sample responded with the *No Opinion* option. This implies mixed or perhaps uncertain opinions about the appropriateness of the FP exit criteria.

To further understand their responses, their grades in the FP assessment were investigated to identify any possible association between their responses and their grades. No evidence of clear significant linear correlation was found; this means that the students' perceptions on lowering the FP exit criteria did not systematically correlate with their grades in the FP assessment. However, a non-linear association could be generally noticed from cross-tabulating the students' responses and grades in Table 8.3 and Figure 8.1 where (1) most of the students who obtained a grade of 2.7 or higher on the FP assessment tended to disagree with this item, (2) most of the students who obtained a grade between 1.3 and 2.6 tended to agree with it, and (3) most of the students who obtained a grade of 1.2 or lower tended to disagree with it. This means that most students who obtained 75% of the total score or more or who obtained 55% of the total score or less on the FP assessment believed that the FP exit criteria should not be lowered to allow more students to join the FY academic study. Although it is hard to offer a certain explanation of such disagreement with lowering the FP exit criteria, it could be speculated that the students with lower grades in the FP assessment perhaps found it challenging to cope with the FY linguistic demands; and that the students with higher grades in the FP assessment disagreed with lowering the linguistic bar of the FY study which, in their minds, might have been linked to "dumbing down" the level of the academic content.

Table 8.3. Cross-tabulation of Student Responses to *Dissatisfaction with Language Assessment* (Item 1.1) with their Grades in the Foundation Programme

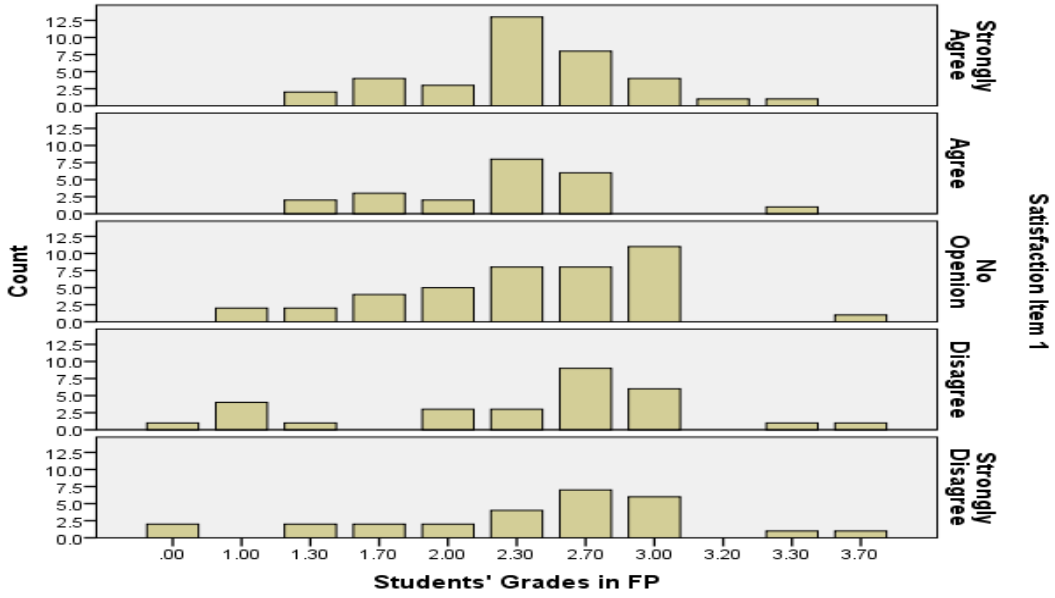
Grades[a] in FP	**Item 1.1:** Assessment on the FP should have allowed more students to proceed to the FY.					**(*N*=155)**
	SA	A	NO	D	SD	
.00[b]	0	0	0	1	2	3
1.00	0	0	2	4	0	6
1.30	2	2	2	1	2	9
1.70	4	3	4	0	2	13
2.00	3	2	5	3	2	15
2.30	13	8	8	3	4	36
2.70	8	6	8	9	7	38
3.00	4	0	11	6	6	27
3.20	1	0	0	0	0	1
3.30	1	1	0	1	1	4
3.70	0	0	1	1	1	3
Total	**36**	**22**	**41**	**29**	**27**	**155**

This table includes only the students who responded to the questionnaire and whose grades in FP assessment were retrievable.

[a] These grades mean the followings in term of scores out of 100.
0<50, 1.00=50 – 54, 1.30= 55-59, 1.70= 60-64, 2.00=65-69, 2.30=70-74, 2.70=75-79, 3.00=80-84, 3.30=85-89, 3.70=90-94, and =95-100. The complete scale is presented in Chapter 10, section 10.2.

[b] Failed in FP assessment in June 2011 but passed FP assessment in August 2011 and joined FY in September 2011

Figure 8.1. Students' Responses on Lowering FP Criteria by their Grades in FP



Second, the students' responses to the items about the social and political impact of English language assessment in this phase indicated similar opinions to those

expressed in Phase 1 of the study presented in Chapter 7. In this phase, a political impact of the FP English language assessment was recognised by most of the participants but not a social one. The results showed that 89.8% of the students seemed to believe that English language assessment was vital for future employment; and 80.6% believed that English language assessment in higher education could have an effect on the country's international status. However, only 32.9% of the students seemed to believe that the FP assessment entailed more negative social consequences than the FY assessment.

8.2.3. Means and Standard Deviations of Students' Responses to the Questionnaire Topics

In this section, the participants' responses to all items under a specific topic were aggregated and averaged as has been explained in Section 6.2.2. The meaning of the responses to the topics are discussed, taking into account the responses to the individual items which sometimes reveal a different aspect of the students' views from that apparent from the general averaged responses. The topic means are shown in the table below, in an ascending order. In this likert scale, means < 3.0 signify agreement, means > 3.0 signify disagreement and a mean of 3.0 signifies *No Opinion* or equal responses of agreement and disagreement to an item.

The lowest mean was that of the *Predictive Validity* topic (M =1.76). This suggests that the majority of the students seemed to believe that their performances in the English language assessment influenced their performances in academic assessment; and that they needed additional English language courses to academically perform better.

The means of the responses to *Assessing Content and Language in English Language Course* and *Assessing Content and Language in Academic Courses* indicate a belief that the content and language accuracy of written assignments in the academic courses should be assessed, but to a lesser degree compared to assessing content and language accuracy in the English language courses, as the means were 2.16 and 2.65 respectively; both means fell in the 'agreement' range of the Likert scale. However, a closer investigation of the items under the two topics revealed that most of the students seemed to agree that the content and language

accuracy of their written expressions should be assessed in both courses (in language courses 89% of the participants expressed agreement and in the academic courses 86%). However, they moderately consented with the view that their teachers actually did assess both the language accuracy and content of written assignments (in language courses 46% of the participants expressed agreement and in academic courses 68% of the participants expressed agreement). The lower percentage of students who agreed with the second view is possibly explained by the students' uncertainty about whether their teachers considered the accuracy of language in assessing written assignments as revealed by the results of focus groups in Chapter 9.

The other two topics with the lowest means are *Impact of FP Assessment* (M=2.25) and *Dissatisfaction with FP Assessment* (M=2.26). These means seem to suggest the majority of students' recognition of the impact of the English language assessment and their general dissatisfaction with FP assessment. However, considering the students' responses to the individual items, a different and more complicated picture emerges. As has been pointed out earlier, the students' responses to *Dissatisfaction with Language Assessment* signalled students' dissatisfaction through their approval for changing FP and FY language assessment, but also revealed a variety of views to whether more students should be allowed to pass the FP assessment. Likewise, their responses to *Impact of FP Assessment* showed that their apparent recognition of the power of the FP language assessment on national and international policies, but not on social aspects. The students' responses to these two topics showed agreement with one aspect of the topic and disagreement with another. Here, the importance of investigating the individual items of the scale, besides the average responses to the general topics is especially obvious

Table 8.4. Means and Standard Deviations of Responses to Student Questionnaire in Phase2

Topic	Minimum	Maximum	Mean	Std. Deviation
FP Assessment Predictive validity	1	5	1.76	.71
Assessing Content and Language in English Courses	1	5	2.16	.69
Assessing Content and Language in Academic Courses	1	4.60	2.22	.61
Impact of FP Assessment	1	5	2.25	.72
Dissatisfaction with FP Assessment	1	5	2.26	.69
First Year Assessment Construct Validity	1	5	3.10	1.03
Adequacy of Language Level for First Year Study	2	5	3.12	.68

The two highest means which indicate disagreement were those of the *FY Construct Validity* (M= 3.22) and *Adequacy of English Language Level* (M=3.09). Most of the students' responses to the items under *FY Assessment Construct Validity* indicated a cautious agreement with the appropriateness of the FY English language assessment instruments, and a moderate disagreement with the view that language accuracy and content were always considered when assessing a students' language proficiency. Similarly, in responding to the *Adequacy of the English Language Levels for FY Study*, the responses to the individual items under this topic varied. Though most of the students seemed to face difficulties in understanding the lectures, reading the assigned texts or expressing their ideas in writing, they strongly felt that their English language levels were adequate for undertaking the FY study. The opinions on this area are somewhat different from the attitudes expressed in student focus groups as will be discussed in Section 9.2.2; most of the students in focus groups seemed to feel that their English language levels were inadequate for undertaking the FY courses. The discrepancy in the responses to the same issue could be due to the nature of questionnaires which imposes specific and limited room for expressing views, unlike focus groups which allow expansion. Also the two methods sometimes generate data that is different not only in type but in content as well.

8.2.4. Comparing Students' Perceptions across the Groups

In order to explore the significance of these differences, a Mann-Whitney U Test was used for the categories that included two groups and a Kruskal-Wallis Test was used for those that included more than two. The results showed significant differences in students' responses across the college, specialization and self-evaluation groups, but not between the gender groups.

As has been explained in Chapter 6, Likert scales produce ordinal or categorical data which is best investigated using non-parametric tests. Also, the data was tested for normality of distribution using Kolmogorov-Smirnov tests, skewness values and histograms. The results showed that the data was not normally distributed (see appendix 8.1), therefore, only non-parametric tests were used to investigate this data set.

8.2.4.1. Differences between College Groups

According to their college groups, students' responses were found to be significantly different in two topics: *Construct Validity* and *Dissatisfaction with FP Assessment*. Sur students (n= 54) seemed to perceive the construct validity of the FY assessment more positively than did Rustaq students (n= 122) U =2.605, Z= -2.23, p =0.25. Figure 8.1 displays the distribution of the students' responses on this topic by their colleges. However, Sur students seemed to be more dissatisfied with the FP assessment than did Rustaq students. The difference between Rustaq and Sur students' responses was significant, U= 5.527, Z=2.54, p= 0.01. The differences between the two colleges in responding to each topic are shown in Figures 8.2 & 8.3 and Table 8.5.

Figure 8.2. Students' Responses to *First Year Assessment Construct Validity* by Colleges

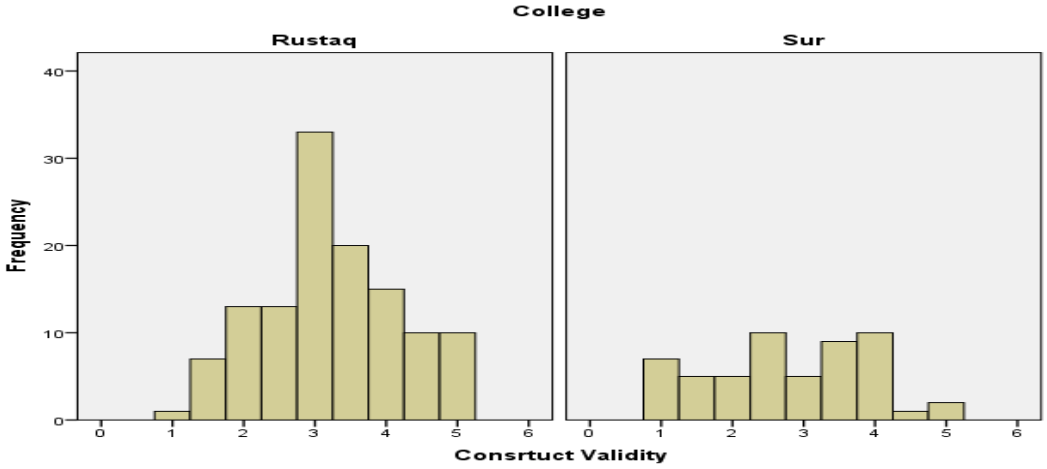


Figure 8.3. Students' Responses to *Dissatisfaction with FP Assessment* by Colleges

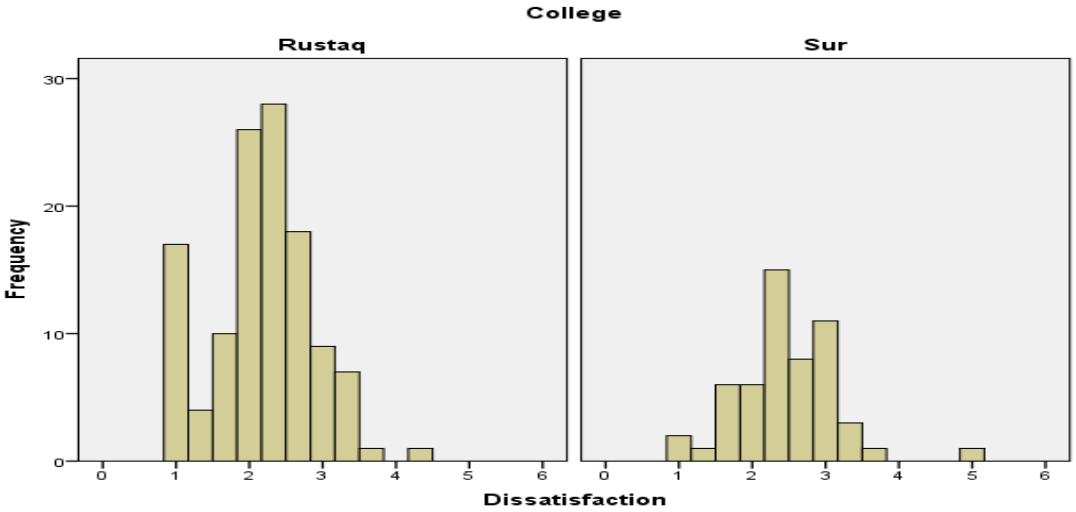


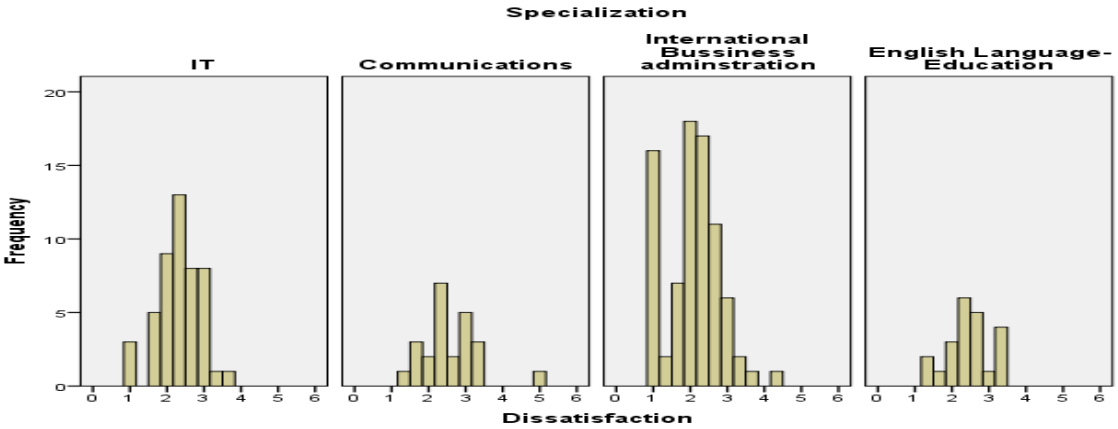
Table 8.5. Means of Student Reponses to Phase 2 Questionnaire by Colleges

Questionnaire Topics	**College**	
	Rustaq (*n*= 122)	**Sur (*n*=45)**
Assessing Language and Content in English Courses	2.17	2.13
Assessing Language and Content in Academic Course	2.65	2.67
Construct Validity	3.23	2.79
Adequacy of Language Level	3.20	3.25
Predictive Validity	1.76	1.78
Dissatisfaction with FP Assessment	2.17	2.47
Impact	2.27	2.19

8.2.4.2. Differences among Specialization Groups

The significance of the differences in the responses to the questionnaire across the four specializations was tested by Kruskal-Wallis. The results showed that there was a significant difference across the specializations in *Dissatisfaction with FP Assessment*, X^2 (3, n=175) = 12.09, p =.007. Table 8.6 below shows that the CS group (M=2.58) and the *English Language* group (M=2.45) were less dissatisfied with FP assessment than were the IT group (M=2.33) and *IBA* group (M=2.01). [To remind the reader, lower means signified more agreement with the items of this questionnaire.]

Figure 8.4. Students' Responses to *Dissatisfaction with FP Assessment* by Specializations



The significant differences in *Dissatisfaction with FP Assessment* that were found among the groupings by specializations could be linked to similar significant differences found in the responses to the same topic between the colleges. As the IBA and IT groups expressed more dissatisfaction than did the other groups and since most of them were from Rustaq College, the significant differences found between the two colleges could be a result of specialization distribution in each College (see Table 8.6).

Table 8.6. Distribution of Students by College and Specialization

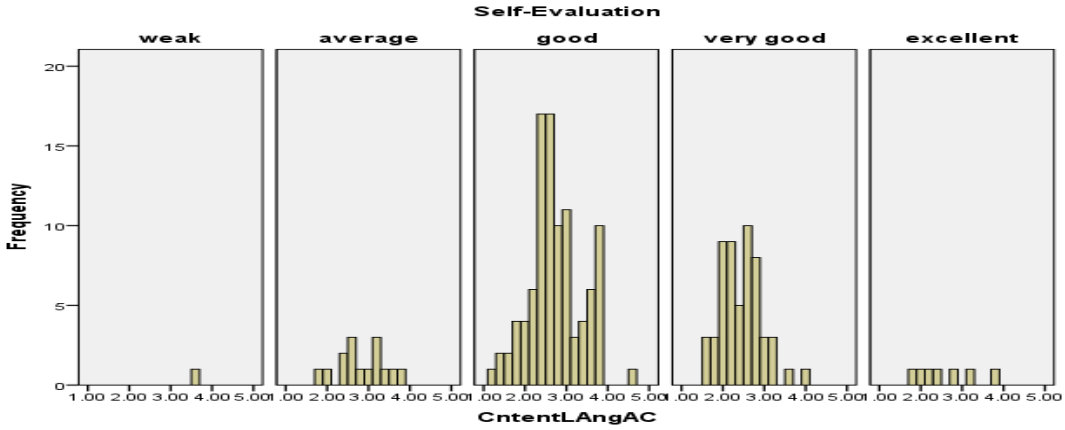
College		***n***	**%**
Rustaq	Information Technology	18	14.2
	International Business Administration	85	66.9
	English Language-Education	24	18.9
	Total	**127**	**100.0**
Sur	Information Technology	32	56.1
	Communication Studies	25	43.9
	Total	**57**	**100.0**

8.2.4.3. Differences among Self-Evaluation Groups

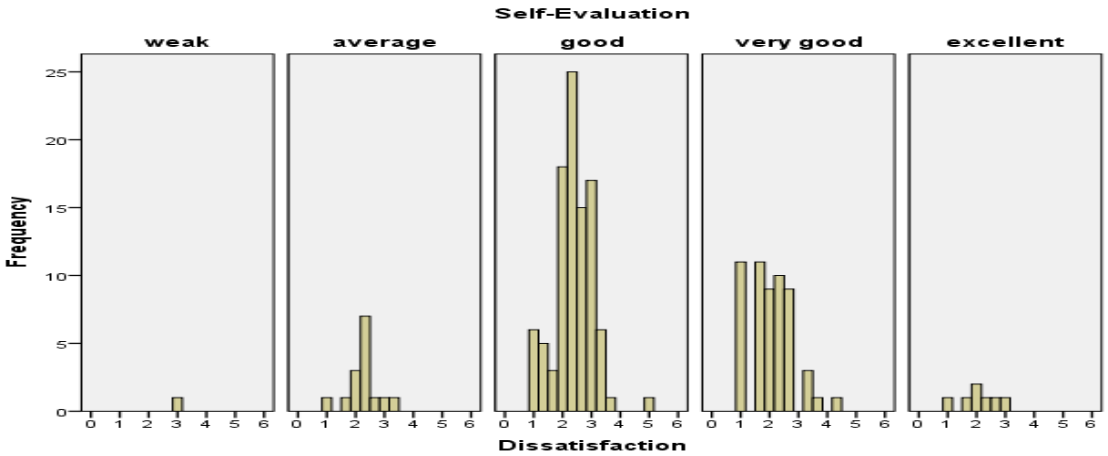
A Kruskal-Wallis test was used to evaluate the significance of these differences. The *weak* group, which included one participant, was combined with the *average* group because this test requires a minimum of five cases; the rest of the groups remained intact. The results revealed significant differences amongst the four groups (Gp1, n=16: *average*, Gp2, n=98, GP3: *good*, n= 55: *very good*, GP4, n=7: *excellent*) in responding to the *Dissatisfaction with FP, Assessing Language and Content in Academic Courses* and *Adequacy of Language Level for FY Study*

topics. For the *Dissatisfaction with FP* topic, the results were X^2 (3, n=175) = 10.42, p= .01. For the *Assessing Language and Content in Academic Courses*, the results were X^2 (3, n=176) = 12.25, p .016. For the *Adequacy of English Language Level for FY Study*, the results were X^2 (3, n=175) = 24.11, p =.01. These differences are apparent in the three figures below.

Graph 8.5. Students' Responses to *Assessing Content and Language* by Self-Evaluations



Graph 8.6. Students' Responses to *Dissatisfaction with FP Assessment* by Self-Evluation



Graph 8.7. Students' Responses to *Adcquacy of Language Levels for FY Study* by Self-Evaluation

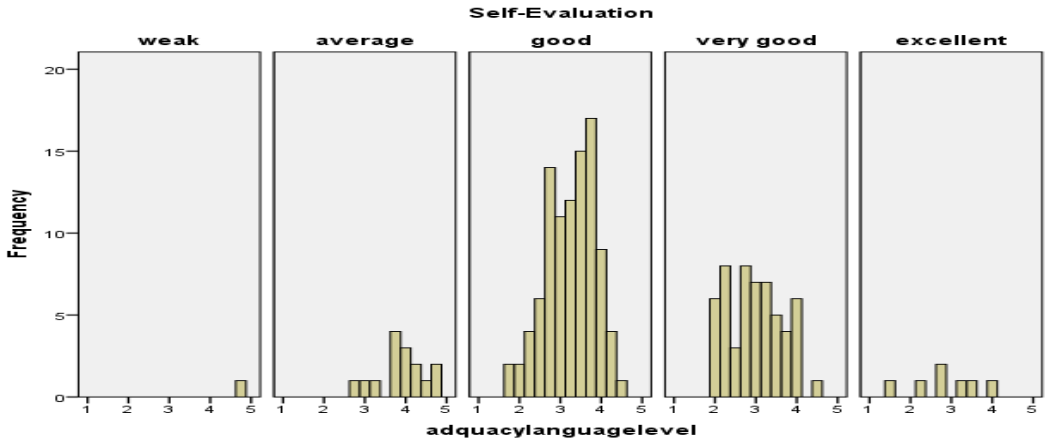


Table 8.8 displays the means of the responses of each self-evaluation group to the three topics mentioned above. The means seem to indicate that (1) the higher the students evaluated their English language levels the less satisfied they tended to be with FP assessment, (2) the lower the students evaluated their language skills the less they tended to opt for assessing both the language and content in academic courses except in the *excellent* group which could be described as being less enthusiastic about this issue than the *very good* group, and (3) with higher levels of self-evaluation, students seemed to be less satisfied with the adequacy of their language skills to meet the FY language requirements. The *excellent* group's responses showed slightly more satisfaction than did the *very good* group's in responding to the same topic.

Table 8.7. Means of Responses to Three Topics by Self-Evaluation

Self-Evaluation	**Dissatisfaction with FP**	**Assessing Language and Content in Academic Courses**	**Adequacy of Language Levels for FY Study**
Average	2.31	2.36	3.95
(*n*=16)	16	16	16
Good	2.39	2.26	3.26
(*n*=98)	98	98	98
Very Good	2.05	2.03	2.99
(*n*=55	55	55	55
Excellent	2.10	2.20	2.86
(*n*=7)	7	7	7
Total	2.26	2.20	3.22
	176	176	176

It should be noted, however, that the correlation between the students' self-evaluations and their FY grades in the English language course was very weak. A Spearman correlation coefficient was used to investigate this relationship, and showed a non-significant correlation rho = - .052, p = .56.

8.3. The Teacher Questionnaire in Phase 2

8.3.1. Demographic Characteristics of the Participants

In the second phase, 29 teachers completed this teacher questionnaire; 14 (48.3%) were from Sur College and 15 (51.7%) from Rustaq College. In the sample, there were 14 female teachers (48.3%) and 15 male teachers (51.7%). Of the participants 9 (31%) were Omanis and 20 (69%) were non-Omanis. In this phase, unlike the first phase, the invited teachers were from different departments including the English language department; the participants' age, education and departments are displayed in the following two tables.

Table 8.8. Classification of Teachers by Age and Department in Phase 2

Age	**Frequency (*N*=29)**	**%**	**Department**	**Frequency (*N*=29)**	**%**
20-30	7	24.1	**CS**	3	10.3
31-40	13	44.8	**IT**	6	20.7
41-50	6	20.7	**IBA**	4	13.8
51-60	3	10.3	**English Language (Education)**	16	55.2

Table 8.9. Frequency and Means of Teachers' Responses to the Teacher Questionnaire in Phase 2

Topic	**Subtopic**	**Items**	**SA 1**	**A 2**	**NO 3**	**D 4**	**SD 5**	**Mean**	**Mean** (Recoded)
Consistency between First Year and Foundation Programme English Language Assessment	___	1.1. In general, assessment of student performance on the FP and in the FY English course provides similar results for the majority of students.	1 (3.4%)	11 (37.9%)	10 (34.5%)	6 (20.7%)	1 (3.4%)	2.83	–
		1.2. There is a close relationship between student performance on the FP and in their performance in the FY English course.	9 (31.0%)	13 (44.8%)	6 (20.7%)	-	1 (3.4%)	2.00	–
		1.3. Assessment should be standardised within CAS.	5 (17.2%)	16 (55.2%)	6 (20.7%)	2 (6.9%)	-	2.17	–
		1.4. There is a close correlation between student performance on the FP and their performance in FY academic courses.	2 (6.9%)	17 (58.6%)	7 (24.1%)	3 (10.3%)	-	2.38	–
Foundation Programme Assessment Validity	**predictive**	2.1. Students do better in FY academic courses when their English language scores on the FP are higher.	15 (51.7%)	11 (37.9%)	2 (6.9%)	1 (3.4%)	-	1.62	–
		2.2. If students perform well in English language courses, they will perform well in First Year academic courses too.	4 (13.8%)	19 (65.5%)	2 (6.9%)	4 (13.8%)	-	2.21	–
		2.3. Students' weak performance in FY academic courses could be caused by factors other than their English language levels. (Recoded)	6 (20.7%)	15 (51.7%)	2 (6.9%)	5 (17.2%)	1 (3.4%)	2.31	3.68
		2.4. The low English language level of some students in the FY causes them to fail FY academic courses.	7 (24.1%)	19 (65.5%)	2 (6.9%)	1 (3.4%)	-	1.9	–
		2.5. Students' language levels influence	8	20	1	-	-	1.76	_

		their achievement in FY academic courses.	(27.6%)	(69.0%)	(3.4%)				
	Construct	2.6. When allowing students to pass into the F Y, it is more informative to focus on students' results in individual English language skills on the FP (e.g. writing, listening or reading marks) than on their total marks.	4 (13.8%)	13 (44.8%)	4 (13.8%)	7 (24.1%)	1 (3.4%)	2.59	–
		2.7. Students' scores in all language skills (reading, writing, speaking and listening) assessment on the FP are equally important indicators of their future academic achievement in the FY.	4 (13.8%)	19 (65.5%)	1 (3.4%)	5 (17.2%)	-	2.24	–
Satisfaction with Assessment	**FP**	3.1. Most students admitted to the FY have the appropriate English language skills to understand and communicate in their academic courses.	1 (3.4%)	9 (31.0%)	6 (20.7%)	11 (37.9%)	2 (6.9%)	3.14	–
		3.2. Students' current language abilities are generally adequate for the academic courses in the FY.	-	11 (37.9%)	6 (20.7%)	10 (34.5%)	2 (6.9%)	3.10	–
		3.6. English language assessment on the FP effectively measured students' abilities to function in FY academic courses.	1 (3.4%)	13 (44.8%)	7 (24.1%)	8 (27.6%)	-	2.76	
	FY	3.7. Assessment instruments in the FY English course focus on the academic language skills students need in FY academic courses.	3 (10.3%)	13 (48.3%)	8 (27.6%)	5 (13.8%)	-	2.45	–
		3.5. FY English course assessment measures students' academic language use efficiently.	-	17 (58.6%)	8 (27.6%)	4 (13.8%)	-	2.55	–

Assessing Language Accuracy in First Year Academic Courses	——	4.1. Assessment criteria in the academic courses should not include students' English language level. (Recoded)	4 (13.8%)	12 (41.4%)	8 (27.6%)	5 (17.2%)	-	2.48	3.51
		4.2. One of the criteria used to mark the FY academic courses should be English language competence	9 (17.2%)	15 (51.7%)	5 (13.0%)	-	-	1.86	–
		4.3. Academic course assessment should aim to be less dependent on students' language ability.(Recoded)	3 (10.3%)	8 (27.6%)	2 (6.9%)	14 (48.3%)	2 (6.9%)	3.14	–
		4.4. When assessing academic courses, markers should overlook language inaccuracies as long as the meaning is clear.(Recoded)	1 (3.4%)	15 (51.7%)	4 (13.8%)	6 (20.7%)	3 (10.3%)	2.83	3.17
		4.5. Teachers in academic courses should assess students on their written expressions as well as their ideas.	8 (27.6%)	16 (55.2%)	3 (10.3%)	1 (3.4%)	1 (3.4%)	2.00	–
Impact of the FY Assessment	**Social Impact**	5.1. The current assessment instruments take account of other parties' opinions (e.g. students). (Recoded)	1 (3.4%)	6 (20.7%)	9 (31.0%)	12 (41.4%)	1 (3.4%)	3.21	2.76
		5.2. Planning how to assess students' work is a process to which teachers, students, society and other related organizations should contribute. (Recoded)	6 (20.7%)	9 (31.0%)	3 (10.3%)	9 (31.0%)	2 (6.9%)	2.72	–
		5.3. In my department, students' opinions about assessment instruments are considered in the design of assessment instruments.(Recoded)	6 (20.7%)	-	8 (27.6%)	11 (37.9%)	4 (13.8%)	3.25	2.75
	Political Impact	5.4. English language assessment should not be a gate-keeper to higher education in Oman. (Recoded)	3 (10.3%)	5 (17.2%)	7 (24.1%)	9 (31.0%)	5 (17.2%)	3.28	2.72
		5.5. Assessing and teaching in English creates more employment opportunities for students.	10 (34.5%)	13 (44.8%)	4 (13.8%)	2 (6.9%)	-	1.97	–

		5.6. Being taught and assessed in English makes Oman an active part of the global village.	11 (37.9%)	10 (34.5%)	6 (20.7%)	2 (6.9%)	-	1.93	–

8.3.2. Teachers' Responses to the Individual Items of the Questionnaire

The responses to the items within four topics showed some differences in teachers' opinion about *Satisfaction with FP Assessment*, *Assessing Language and Content in Academic Courses*, and *Social Impact*. In the first topic, though most of the teachers seemed dissatisfied with the appropriateness of students' English language levels for FY courses, as the responses to items 3.1 and 3.2 indicated, they seemed to feel that FP assessment was a good measure of the students' language abilities in item 3.3. It seems that the teachers were satisfied with the assessment instruments themselves, but not with the students' levels and they tended to differentiate in their responses between the two.

Another topic that produced superficially contradictory opinions was the items on assessing the language accuracy of written assignments in the academic courses. Most of the teachers agreed that language should be a criterion in assessing written assignments in the academic courses (item 4.2 and item 4.5), but felt that language inaccuracies should be overlooked when the intended meaning is comprehensible (item 4.4).

8.3.3. Means and Standard Deviations for the Questionnaire Topics

Table 8.10 displays the means of teachers' responses to each of the questionnaire topics in an ascending order.

Table 8.10. Means and Standard Deviations of the Responses to Teacher Questionnaire Topics

Topics	**Min.**	**Max.**	**Mean**	**Std. Deviation**
Political Impact	1.00	4.00	2.20	.77
FP Predictive Validity	1.20	3.20	2.22	.52
Consistency in FY and FP Assessment	1.50	3.75	2.31	.51
FP Construct Validity	1.00	4.00	2.44	.74
FY Satisfaction	1.50	3.50	2.50	.56
Assessing Language in FY Academic Courses	1.80	4.00	2.68	.53
Social Impact	1.00	4.67	2.74	.75
FP Satisfaction	1.67	4.67	3	.76

The lowest mean was of the *FY Political Impact* topic (M= 2.2) and the highest mean was of the *FP Satisfaction* topic (M= 3.0). This indicates that the teachers viewed English language assessment in FY as having political impact, and that

they had mixed feelings about FP assessment. Many teachers seemed to be more satisfied with FY English language assessment (Mean=2.87) than they were with FP English language assessment (Mean= 3.0). Though their responses to some items on FP validity reflected positive perceptions (means were less than 2.4), their overall responses to the items on FP satisfaction tended to be less positive.

Likewise, the teachers seemed to believe that there was a positive correlation between the students' scores in the English language course and their scores in the academic courses. The *FP Predictive Validity* Mean value was (M=2.2) which signalled a majority agreement that students' performance in English language assessment could predict their achievement in academic courses assessment.

In line with the results of the first phase, most of the teachers in the second phase believed that the political impact of FY English language assessment was greater than was its social impact. The average Mean value for *Political Impact* was (2.2), and for the Social Impact it was (2.74).

8.3.4. Comparing Teachers' Perceptions amongst the Groups

This section looks into the opinions of the groups within the teacher sample (i.e., gender, department, nationality, age, and teaching and writing assessment experience) using Mann-Whitney U test and Kruskal-Wallis. These non-parametric tests were used because Likert scales normally produce categorical or ordinal data and normality tests showed that the data set was not normally distributed (see appendix 8.4). Pallant (2007) asserted that it is unusual to obtain a normal distribution with social sciences measures, especially with small samples.

8.3.4.1. Differences among the Groupings by College and Gender

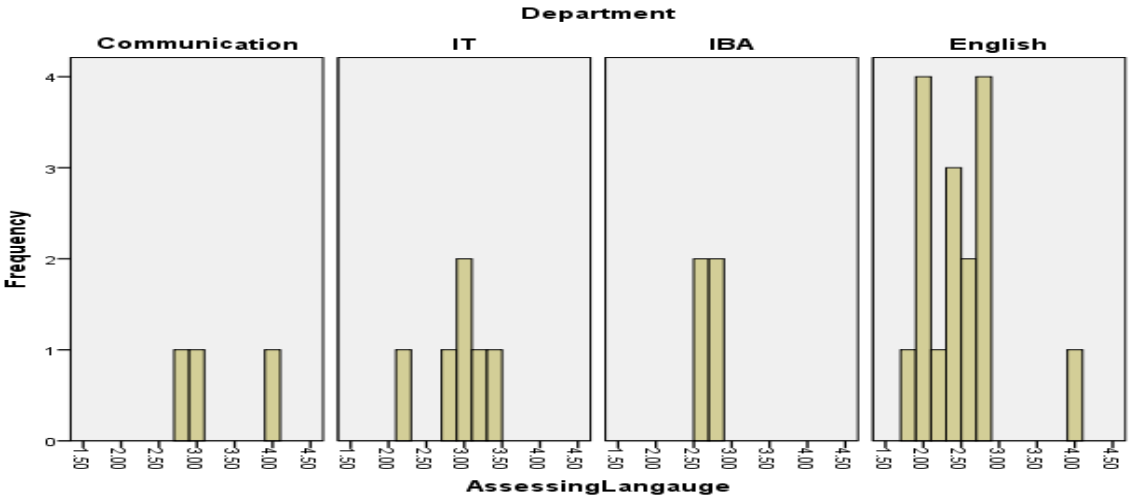
The results revealed that there were no significant differences in teachers' responses to the questionnaire topics according to their college or gender groups.

8.3.4.2. Differences among the Department Groups

In contrast, the teachers' views analysed by their departments (i.e., CS, IBA, IT or English Language-Education) showed substantial differences in responding to seven

topics: *FP Predictive Validity, Consistency in FP and FY Assessment, FP Construct Validity, FP Satisfaction, FY Satisfaction, Assessing language in FY Academic Courses, Social* and *Political Impact*. A Kruskal-Wallis test was used to investigate the significance of these differences. There was a significant difference amongst the teachers' responses to *Assessing Language in FY Academic Courses* only. The groups (Gp1, n= 3: CS), (Gp2, n= 6: IT), (Gp3, n= 4: IBA) and (Gp4, n=16: English) were significantly different in their responses, X^2 (3, n=29) = 9.91, p=.01. The CS group reached a higher mean (M=3.26) than the other three groups.

Figure 8.8. Teachers' Responses to Assessing Language Accuracy in Academic Courses by Departments



This indicates that many of the CS teachers seemed to disagree with the idea that language should be assessed in academic courses, whereas, the teachers from the other departments expressed various degrees of agreement with using English as a criterion in assessing academic written assignments. Figure 8.8 illustrates the differences in responding to the topic *Assessing Language in FY Academic Courses* amongst the groupings by department.

8.4. Discussion

In this section, the results from both questionnaires will be compared to identify and try to explain the differences in the teachers' and students' responses, discuss the similarities in their responses, and identify the significant differences in

responses across the groups. It will also link the results to related literature in an attempt to explain, clarify or sometimes consider counter-arguments to the ones discussed in this chapter. In some places, the findings will be compared to those of other chapters to build a coherent picture.

8.4.1. The Students' and Teachers' Responses

The results from the student and teacher questionnaires showed similarities and differences in their views about the *Satisfaction with FP Assessment, FP Predictive Validity, FY Assessment, Assessing Language Accuracy in FY Academic Courses,* and *Impact of FY English Language Assessment*. These views are encapsulated in the means of the teachers' and students' responses to the questionnaire topics presented in an ascending order in the table below.

Table 8.11. Means of Responses to Teacher and Student Questionnaires in Ascending Order

Teacher Questionnaire Topics	**Mean**	**Student Questionnaire Topics**	**Mean**
Political Impact	2.20	FP Predictive Validity	1.76
FP Predictive Validity	2.22	Assessing Content and Language in English Courses	2.22
Consistency in FY and FP Assessment	2.31	Impact	2.25
FP Construct Validity	2.44	Dissatisfaction with FP assessment	2.26
FY Satisfaction	2.50	Assessing Content and Language in Academic Courses	2.65
Assessing Language in FY Academic Courses	2.68	FY Construct Validity	3.10
Social Impact	2.87	Adequacy of English language level for FY study	3.12
FP Satisfaction	3		

The students' and teachers' views differed in two topics: *Satisfaction with FP Assessment* and *Validity of FP Assessment.* The students' responses seemed to indicate a moderate dissatisfaction with FP English language assessment (M=2.26), whereas their teachers' responses showed mixed feelings about FP assessment (M=3.0). The students' dissatisfaction fits in with their negative responses to the items on the adequacy of their English language levels for FY study (i.e, items 2.2, 2.3, and 2.4). They seemed to believe that the difficulties faced in FY are a result of their inadequate English language levels. In contrast, the teachers' responses reflected a positive view of the predictive and construct aspects of FP assessment validity with means of (2.22) and (2.44). Also the teachers, unlike the students,

expressed satisfaction with FY English language assessment with a mean of (M=2.50). These contrary perceptions of the appropriateness of FP and FY assessment could be understood in the light of findings presented in a conference paper on the IELTS predictive validity of academic achievement (Bayliss, 2006). The researcher reported that the students seemed to believe that their language skills were less adequate than did their teachers. In line with Bayliss's findings, the results of the questionnaires suggest that most of the students seemed to perceive their language levels as inadequate more than did their teachers. However, this preliminary suggestion is challenged by the results obtained from the focus groups and interviews presented in Chapter 9 in which the teachers and students equally seemed to believe that the students' language levels were inadequate for FY study (see Sections 9.2.2. and 9.3.2).

Though the teachers' and the students' responses to the questionnaires reflected disagreement about their satisfaction levels of FP assessment, they both seemed to agree that English language proficiency predicts achievement in academic courses. The mean of the teachers' responses to the *FP Predictive Validity* was (M= 2.2), and the mean of the students' responses to the same topic was (M=1.7). It seems that the students believed that performance in English language positively correlated with performance in academic courses more than did their teachers. Though only a few studies, such as Powers, Kim and Weng's (2008), have investigated students' and teachers' perceptions of the predictive validity of English language assessment, a plethora of studies have investigated the strength of the predictive validity of the English language assessment and mostly reported low predictive validity values (e.g., Elder, 1993; Lynch, 2000; Davies, 2009).

Likewise, both the teachers and students seem to think that English language should be used as a criterion in assessing students' written assignments in academic courses. However, their responses to the individual items under this topic suggested different attitudes towards this issue. In the student questionnaire, most of the respondents seem to support assessing language in academic courses, but were not certain whether their teachers did in fact mark language in the academic assignments or not. In the teacher questionnaire, though many teachers thought that language should be

assessed in academic written assignments, most of them seemed to believe that it should be overlooked if the content of a written piece was comprehensible. The teachers' "double standard" on this matter and the students' views are further investigated and discussed in Chapter 9 (see Sections 9.2.4 and 9.3.4).

Similarly, the impact of English language assessment seemed to be recognised by both the students and teachers. The majority of the students agreed with the statements that the English language played an important role in future careers and the international position of the country (M=2.25). Likewise, the political impact of the English language assessment was recognised by most of the teachers (M=2.2), who also seemed to think that English language assessment had a social impact (M=2.68), but to a lesser degree than the political one. This difference was also apparent in the first phase results where both the students and the teachers seemed to recognise the political impact but not the social one (see Section 6.4.2). This finding can perhaps be understood within the wider picture painted by Shohamy (2006) where English language plays a significant role in the global market and tests are used internationally to enforce mostly covert political agendas as has been discussed in Chapter 1 (see Section1.3.2).

8.4.2. Significant Differences among the Groups in the Teacher and Student Questionnaires

Only few topics in the two questionnaires showed significant differences amongst the groups. The differences in the groups' responses to *Assessing Language in Academic Courses* were significant in both the teacher and student questionnaires. The self-evaluation groups of the students responded differently to this topic. The students, with the lower levels of self-evaluation, were less convinced that English language should be assessed in academic courses. The specialization groups of the teachers also responded differently to this topic. The English language teachers (M= 2.4) agreed more with this view than did the IBA teachers (M=2.7) and the IT teachers (M=2.9). On the other hand, the CS teachers showed on average a slight disagreement with this view (M=3.2). This indicates that students' views on

assessing language in academic assignments differed according to their self-evaluated language levels while their teachers' views on the same matter differed according to their departments.

All of the other significant differences in responding to the questionnaires were found amongst the student groups' responses. First, their responses to the *Dissatisfaction with FP Assessment* topic were significantly different amongst the groupings by college, specialization and self-evaluation. Sur students were more dissatisfied with FP assessment than were Rustaq students. Also, the IBA students were more dissatisfied with FP assessment than were the IT students who were more dissatisfied with FP assessment than the English language students. The CS students were the least dissatisfied. As explained earlier, the significant difference in Sur and Rustaq students' responses perhaps could be a result of the sample distribution in each specialization per college. The items about the satisfaction with the FP Assessment were also responded to differently amongst the student self-evaluation groups. The *very good* group seemed to be the most dissatisfied by FP assessment followed by the *excellent* group, the *average* group and the *good* group respectively.

Second, students' responses to *FY construct Validity* showed significant differences amongst the college groups; and their responses to the *Adequacy of their language levels for FY study* showed significant differences across the self-evaluation groups. Most of the Sur students tended to express moderate agreement with the items on the *FY Construct Validity*, however, Rustaq students tended to express disagreement with the same items. On the *Adequacy of Language Levels for FY Study*, the lower the students' self-evaluation levels were the more they disagreed with the items of this topic. Though several studies reported moderate to strong correlations between self-evaluation and performance in formal language assessment (e.g., Blanche, 1989; Ross, 1998), this study found a non-significant correlation. Therefore, the reported significant differences between the self-evaluation groups should be understood as students' perceptions of their language proficiency that do not reflect their language proficiency as measured by formal assessment instruments.

8.5. Summary and Concluding Remarks

In general, the questionnaires revealed more similar views than different between the students' and teachers' responses. Though the students were less satisfied with the adequacy of the language levels than were their teachers, they both strongly agreed on the predictive validity and the political impact of English language assessment. Also, the CS group in the student and teacher samples showed significant differences from other groups on their views on assessing language in academic courses (in the teacher sample) and dissatisfaction with FP assessment (in the student sample).

The findings of this questionnaire confirmed some of the findings presented in Chapters 6 and 7. In previous Chapters, as is in this one, most students and most teachers seemed to believe that proficiency in English language as measured by scores in FP assessment was strongly associated with academic achievement as measured by scores in academic courses assessment. Also, the finding that CS teachers seemed less keen on considering language accuracy while marking academic written assignments explain some of the findings on document analysis presented in Chapter 5. Throughout this thesis, the findings suggest that there is a lack of clarity on assessment requirements not only on the students' part but also for the teachers'; This chapter revealed students' and teachers' confusion about whether language accuracy should be /was used as a criterion in marking academic assignments. More discussion of the implications of these findings is in Chapter 11.

Chapter 9: Results from Student Focus Groups and Teacher Interviews in Phase 2

9.1. Introduction:

This chapter presents the results obtained from the student focus groups and teacher interviews conducted in the second phase of this study. The purpose of collecting data using these two methods was to answer the study questions displayed in the box below.

Box 9.1. The Study Questions addressed by the Focus Groups and Interviews in Phase 2

4. How did the stakeholders[15] understand the relationship between the student performance in the English language assessment and their performance in the academic courses assessment?

> 4.1. What were student and teacher perceptions of issues related to the design, marking and impact of the English language assessment?
>
> 4.2. How did student and teacher think language accuracy should be considered in assessing academic assignments?
>
> 4.3. What were the student and teacher perceptions about the importance of the predictive validity?

2.2. What were the student and teacher perceptions of the assessment tools' effectiveness and their roles in shaping language assessment in retrospect?

The chapter is divided into three main sections. The first reports on the results from the student focus groups categorised into five themes: perceptions on the Foundation Programme (FP) predictive validity, difficulties faced in the First Year[16] (FY) study, issues with assessment tasks and implementations, how language accuracy is assessed in the FY, and the FP in retrospect. The second section presents the results of the teacher interviews organised under five similar themes. The third section brings together the results from the focus groups and interviews under four general topics: relationship between language proficiency and academic achievement,

[15] The term *Stakeholders* refers to teachers and students in this study (see Section 3.4.4.3, for the rationale for this use).
[16] The First Year, despite its name, comes after the Foundation Programme.

language related difficulties in the FY study, effectiveness of FP assessment in retrospect, and assessing language accuracy in the FY written assignments.

9.2. Student Focus Groups

In this section, the results of 15 focus groups conducted in the second phase of the main study are presented according to the themes that emerged when analysing the scripts. The focus groups were conducted in the students' first language, Arabic. In this phase, 83 students participated in the focus groups. The numbers, colleges, genders, and specializations of the participants in each of the groups are displayed in Table 9.1. The video-taped focus group discussions were translated to and transcribed in English. In the transcribed discussions, most of the hesitations, sighs, incomplete sentences or non-linguistic expressions were omitted because the analysis process focused on common themes not on language featured similarities/differences of the discussions as has been mentioned in Chapters 4 and 7. Following thematic analysis, "an emphasis is on what is said rather than on how it is said" (Bryman, 2004, p.412).

Table 9.1. Group, College, Gender, and Number of Students in Phase 2 Focus Groups

Group	**College**	**Gender**	***n***	**Specialization**
Group1	Sur	F	4	Communication Studies
Group2	Sur	F	3	Communication Studies
Group3	Sur	M	3	Information Technology
Group4	Sur	M	4	Communication Studies
Group5	Sur	M	6	Communication Studies
Group6	Sur	M	3	Communication Studies
Group7	Sur	M	9	Communication Studies
Group8	Sur	M	3	Information Technology
Group9	Rustaq	F	3	Information Technology
Group10	Rustaq	M	9	International Business Administration & Information Technology
Group11	Rustaq	F	4	International Business Administration
Group 12	Rustaq	M	7	English language- Education
Group13	Rustaq	M	13	English language-Education
Group14	Rustaq	F	5	International Business Administration
Group15	Rustaq	M	5	Information Technology
Total (*N*)			**83**	

The transcribed discussions of the focus groups were studied and 11 codes were created based on the themes that emerged (see Section 4.7.2 for a detailed account of the analysis process). The findings were organised under the five following themes.

9.2.1. Students' Perceptions of the FP Predictive Validity

In the focus groups, the students were asked to discuss the extent to which they believed that their English language proficiency influenced their performance and scores in the English language medium academic courses. In Sur College, all focus group discussions maintained that there was a positive correlation between the students' English language levels and their performance in academic courses. Most of the students seemed to believe that better English language skills resulted in better performance and scores in the academic courses. Some of the discussions on this issue went as follows:

Group (3)

Student 1: Of course, proficiency in English influences scores.
Student 3: You can communicate better with your teacher and tell him anything you want or ask about anything if your language is good. If your language is weak, you cannot participate in class.
Student 3: It is true that sometimes it depends on the mental abilities of the students but if you do not understand what you read in a test, you won't be able to perform well. Also, most courses need a lot of translation, I mean, the textbooks.
Student 1: In mathematics, we solve problems well [in class] but in the exam we understand the problems but we do not understand the question and [we do not understand] whether the teacher wants us to solve it according a certain method or not ... This is all because we did not study well in the foundation and we were not equipped with the right English language skills there.

Group (4)

Student 1: English language affects performance in academic courses because they are taught in English.
Student 4: If your English language level is weak, you cannot study the courses. Communication Studies course depends a lot on the language and you cannot understand it or do well in it if your English is not good.

In Rustaq College, however, the students had various opinions about the effect that proficiency in English had on academic achievement. In eight instances, the students said that students' proficiency in the English language influenced their academic achievement, but sometimes other factors played a more important role in influencing academic achievement. In two other instances, the view that emerged was that the language proficiency had no role in students' academic achievement or scores; achievement in academic courses depended on non-linguistic factors such as: intelligence and ability to memorise large chunks of information. In one instance, the

students argued that proficiency in English played the *major* role in influencing their scores in academic courses. These views are encompassed by the extracts presented below.

Group (10)

Three students (simultaneously): Sure, our scores [in academic courses] are influenced by the scores in the English course…

Student 5: Usually with the theoretical courses we get less scores and this is true for the English course and Business course. But most of the courses are theoretical not practical, so English is very important.

Student 9: For example, a teacher asks you to write a 1000 words essay in the academic courses, so how can we do it without having good English language?

Group (11)

Student 2: It [proficiency in English] does not influence performance in the other courses, but it influences the scores we get in the essays and presentations. The scores we get in the tests are not influenced by the language abilities.

Student 1: In the tests students depend more on their memorization, thinking skills, and writing abilities, but the language skills help you a lot in your presentations, and knowing vocabulary, and help you in writing essays. Having good language skills saves a lot of time in writing essays or studying the textbooks if your English language skills are good.

Some perceived factors that minimised or maximised the role played by English language proficiency in academic achievement can be inferred from the students' discussions presented above. These included how much the assessment instruments required mastery of productive language skills (e.g., presenting or writing), how much the assessment instruments focused on memorization or practical work (e.g., designing a web page), and how much of assigned reading materials the students had to translate before being able to understand the content.

9.2.2. Language Difficulties in the First Year

In both Sur and Rustaq Colleges, all discussions about the language requirements of the FY academic courses concluded that the level of the language of these courses was too difficult for the students and that almost all of the students faced language related obstacles. The difficulties that the students mentioned ranged from using online websites to translate the reading materials for a course to inability to cope with the length and complexity of writing academic reports. In general, most of the students seemed to believe that FY materials are linguistically challenging. Though in five separate instances, individual students from Sur College expressed that the FY

courses did not involve any language difficulty and if some did, this was only at the beginning of the semester. The following extracts are four samples of the thirty discussions in which students maintained that they faced linguistic difficulties in the FY courses.

Group (2)

Student 2: It is also the language of the textbooks which is difficult and when I put them [reading materials] in Wafi [a Translation software], they do not make any sense at all. So what we do as students is just memorise what we are given even if we do not understand it.

Student 1: It is always difficult when studying for an exam because it is always only memorization, and the teachers can tell from the exams because we write incomplete sentences when we forget what we have memorised. And it is very difficult to memorise without understanding.

Group (3)

Student 1: In the foundation, we should have been introduced to Mathematics, IT, and Communication related vocabulary, because now we study but we do not understand the textbooks very well. What we studied in the Foundation Programme was very general … for example, in the communication I feel I have to translate the whole book, it is depressing

Student 2: As he said, the English language courses are in one end and the academic courses are in the other end. The vocabulary items we need are not in the English course.

Group (10)

Student 2: The First Year is difficult.

Student 1: The difficulty in the First Year is in the large amount of the new vocabulary that we have to learn.

Group 11

Student 2: It [FY] is difficult. Some older students told me that this year would be very difficult and needed a lot of effort in studying; there is a lot of vocabulary. So I find it really difficult.

Student 4: We are in shock

Student 3: We are still in shock.

Student 4: We were [in the Foundation Programme] taking things lightly,[and considered] classes for playing and laughing but now, everything is serious and we spend a lot of time studying and translating the vocabulary.

Student 1: Teachers now are very strict in how we write long essays, use references and always warn us of plagiarism. We did not learn about any of this last year, we did not study about the procedures of writing academic essays.

From the previous extracts, it could be noticed that students of both genders and different specializations equally raised similar concerns about the language difficulties they faced in FY academic courses. These difficulties were perceived to be caused by the large amount of novel vocabulary in academic courses, reading large amounts of academic materials and writing extended (i.e., up to 1000 words)

academic essays. One of the methods students followed to overcome these difficulties was translating big chunks of reading materials via online translation sites and plagiarising in writing academic essays; the latter is further discussed in Section 9.2.4 of this chapter.

9.2.3. Issues with Assessment Activities and Implementation

In focus group discussions, some concerns were raised about how the FY teachers implemented assessment tasks. Four issues recurred throughout the discussions. First, the assessment activities plan and deadlines seemed to be unclear to most of the students. For example, it was claimed that some teachers asked them to write a report a week before a given deadline and without a prior notice. A second issue was the opaque criteria used in marking written assignments and presentations. The third issue, which almost all of the groups raised, was the lack of appropriate feedback received on the students' performances. The fourth was about the low number of assessment instruments undertaken during the semester. These four issues are exemplified in the extracts below.

Group (3)

Student 3: No one tells you how you will be evaluated in the English course. We only know that the final test is awarded 50% of the total mark and the coursework is 50% of the total mark but we do not know the details of how we will be given scores, or the criteria that will be used to assess us.

Student 2: I asked yesterday my teacher about the scores distribution. She told me that there is 50% for the final exam and 30% for the project and presentation. I asked her where the other 20% is and she replied that she did not know.

Group (9)

Student 3: It depends on the course itself, sometimes it is the project, exams or quizzes that determine, but often I feel it is not fair because in the presentations he [the teacher] gives the scores for unclear reason, only because he likes the topic or the person.

Student 1: The new teachers do not know students well or their styles, sometimes they are very strict in terms of the presentation timing or speed of delivery…

Student 2: It is true; we should have been told about how we will be evaluated in a presentation and given a list of things that will be evaluated.

Generally, the IBA and IT groups tended to raise the issues implied in the extracts above more than did the CS or English language (education) groups.

In both colleges, the English language teachers were criticised for not marking students' assignments. Furthermore, in Sur College, the CS students mentioned that their academic course teachers were replaced several times during that semester, and that some teachers had strong incomprehensible accents. No such complaints about the academic courses' teachers were conveyed by Rustaq students.

Group (6)

Student 3: I dropped the communication course because it was very condensed and I did not understand anything from the teacher, so I did not want to risk taking it. The teacher is Indian and his accent is very strong and sometimes he uses some Urdu [phrases] too.

Student 1: The other day he said "hey mat-lab"? Meaning "do you understand" And we told him that we find a difficulty to understand your English let alone your Urdu.

Student 2: What happens in the communication course is that we had three teachers during one course. We had an Egyptian, then and Indian and now another Indian teacher, how do they expect us to learn from three different teachers and not get confused?

In this section, five problematic issues were identified concerning how teachers implemented assessment in both academic and English language courses. Some of these issues concerned the implementation of the assessment tasks themselves while others concerned the teachers' teaching styles.

9.2.4. How Language Accuracy was Assessed in the First Year Courses

According to most focus group discussions, it seemed that many students did not have clear instructions on whether language accuracy was considered in marking the academic written assignments or not. In 20 instances, the students affirmed that language systems (e.g., grammar and vocabulary) of written compositions were not only evaluated in the English language courses but also in the academic courses. In nine other instances, the discussions suggested that it was widely believed that the language accuracy was irrelevant in marking academic written assignments. Regardless of the students' beliefs of how language accuracy was considered, in almost all of the focus group discussions on this issue, a lack of clarity of the marking criteria was indicated as the following extracts show.

Group (15)

Student 2: I do not know for sure, but I suppose they [academic courses teachers] focus on ideas, they should not mark the language but the content of the answers.

Student 1: In the communication course, I know that the teacher scores us down on incorrect spellings, but I do not know about the IT course teacher. We have not been informed yet.

Student 4: I think if the sentences are comprehensible, and the ideas are clear, teachers should not consider the language in marking. But if the ideas are not clear, then the teachers can mark us down.

Group (2)

Student 1: In the communication course the teacher focuses a lot on the grammar mistakes, he underlined every incorrect grammar point. It is unlike the English language course were we are supposed to be corrected in language…

Student 2: But the IT course assignments are marked for the ideas only not the language. The teacher asks for the ideas only.

Group (11)

Student 3: It [marking written assignments] depends on the vocabulary you use, organizing the ideas, and other things. This is in all courses.

Student 1: In the exam the teachers focus on checking spelling but not that much, if a letter or two are wrong, they consider the answer to be right regardless.

Interestingly, the majority of the students seemed to believe that plagiarism and essay length were the most important criteria that the teachers focused on when marking written assignments. In the focus groups, the students repeatedly mentioned avoiding "plagiarism" as a criterion in marking written assignments and occasionally seemed to consider it as the core criterion, especially in the English language course.

Group (10)

Student 7: In the English course teachers evaluate the essays on the grammar and vocabulary. But the essays in the IT course for example the teacher looks at the information and ideas students bring forward.

Student 8: What they care about is the cheating percentage from SafeAssign[17] and that the assignment has all of the ideas needed and that students responded correctly to the question.

Group (7)

Student 6: When the essays are uploaded in blackboard [referring to the SafeAssign software] and the reports come with a percentage of plagiarism, teachers take the reports seriously but they do not think about the fact that hundreds of students are writing about similar topics and of

[17] CAS English language departments use SafeAssign to detect plagiarised texts. It is software that finds similarities between any uploaded text and an archive of online data, books, journals… etc. It also assists detecting any cross-colleges plagiarism where students from different colleges submit essays on the same topic.

course the essays will be similar in terms of words and phrases so they should be careful about taking the reports as they are.

9.2.5. Evaluating the Effectiveness of the Foundation Programme Assessment in Retrospect

The students were asked to reflect on the effectiveness of the FP and its assessment in light of their following experiences (i.e., studying in the FY). Almost all of the students from both colleges argued that the foundation courses and assessment should have been "more intensive", "more challenging", and "stricter". It was explained that the students' expectations of studying at a higher education institute were let down by its overly simple curriculum. As many students stated, when they had been admitted to the colleges, they expected the study at the FP to be more challenging and stimulating but they were disappointed when told by sophomore students that the FP was merely a waste of time. It seemed also that the pass/fail criteria encouraged a lazy and careless behaviour towards studying at FP as expressed in the following discussions.

Group (11)

Student 1: I was disappointed when I entered the Foundation Programme because we studied only two courses. We expected more than that. In the other colleges, students get separate courses for the writing skills, speaking skills, listening skills, reading skills and other courses. This was good for students to get more intensive study and courses.

Student 2: There was not enough progress. I expected myself to study more and to put more effort into studying the language. But it was an easy year and we spent it having fun but not really studying on the language.

Group (12)

Student 5: We were taught really easy things and we are now stunned by the English language requirements of the first year.

Student 3: The level of the language skills we studied last year was similar to what we had studied in high school. Everything is shallow and not deep.

Group (8)

Student3: They [the teachers] should have taught us everything in the foundation programme and not let us pass until they knew we will do well in the first year.

Student2: We needed more materials in the Foundation Programme.

Student3: The Foundation Programme should have been streamed based on specialization and more specialization related vocabulary should have been added.

9.3. The Results of the Teacher interviews

This section reports on the common themes that occurred in the 23 teacher interviews conducted in the second phase of the study. The aim of the interviews was mainly to understand both English language and academic courses teachers' perceptions of the appropriateness of the students' language skills for FY study and predictive validity of the FP assessment. Teachers from four different departments (i.e., English language, Communication Studies (CS), International Business Administration (IBA) and Information Technology (IT)) were interviewed. Table 9.2 below gives more information about the participants.

Table 9.2. College, Gender, Nationality and Department of teachers in Phase 2 Interviews

Number	College	Gender	Nationality	Department
Teacher 1	Sur	M	Canadian	English language (EAP)[18]
Teacher 2	Sur	M	Omani	CS
Teacher 3	Sur	M	Omani	IT
Teacher 4	Sur	M	American	English language (EAP)
Teacher 5	Sur	M	British	English language (EAP)
Teacher 6	Sur	M	British	English language (EAP)
Teacher 7	Sur	M	American	English language (EAP)
Teacher 8	Sur	M	Indian	CS
Teacher 9	Sur	F	Swedish	English language (EAP)
Teacher 10	Sur	F	Indian	IT
Teacher 11	Sur	F	Indian	CS
Teacher 12	Rustaq	M	American	English language (EAP)
Teacher 13	Rustaq	M	American	English language (Major)
Teacher 14	Rustaq	M	British	English language (EAP)
Teacher 15	Rustaq	M	American	English language (Major)
Teacher 16	Rustaq	F	Omani	English language (EAP)
Teacher 17	Rustaq	F	Omani	IT
Teacher 18	Rustaq	F	Omani	Business
Teacher 19	Rustaq	F	Omani	CS
Teacher 20	Rustaq	F	Australian	English language (EAP)
Teacher 21	Rustaq	F	British	English language (Major)
Teacher 22	Rustaq	F	Pakistani	CS
Teacher 23	Rustaq	M	British	English language (Major)

The transcribed interviews were read to identify common issues that constructed the primary list of subtopics. The subtopics were filtered to correspond with the study's focus and questions (see Section 4.7.2 for the detailed procedures used to analyse the interviews). Five common themes emerged from the 23 interviews discussed below.

9.3.1. Correlation between English Language Proficiency and Academic Achievement

All interviewed teachers in the English Language Department agreed that there was a positive relationship between the students' English language proficiency and their academic achievement. They explained that, since CAS were English language medium colleges in which all of the textbooks, lectures and tests were delivered in

[18] EAP stands for English for Academic Purposes course. This course is offered for students from the IT, IBA and CS departments. The *English Language (major)* course is offered to students from the Education department who are studying to Teach English for Speakers of Other Languages (TESOL).

English (except for very few courses), the students with better proficiency in English had an advantage in performing well in the academic courses. The more competent the students were in English language, the better their performance would be in the academic courses. This perception is exemplified in the extracts from the interviews below:

> **Teacher 16**: In my case, the students' major is English language itself, the language has a great impact. Those students who have a good level of mastery of the language do well in the courses.
>
> **Teacher 21**: As their academic courses are all in English, I think that having a high level of performance in the academic course, helps if they [students] can read and if they have extensive vocabulary and particular academic words, it will help because the textbook that they are using in the academic courses are all written in English and utilise a lot of academic words.
>
> **Teacher13**: They [proficiency in English & academic achievement] are definitely correlated, but whether it is a causation or not, I do not know.

Seven of the nine teachers from the IT, CS and IBA departments expressed the belief that high levels English had a major role in achieving well in academic courses.

> **Teachers 3:** I find that some students who are good in English or have excellent English language skills are doing well in the IT course. For example, they do not have a problem paraphrasing their answers and they do not copy exact points from the slides used in the class when answering an IT exam.
>
> **Teacher 22:** I was quite upset with the standard of English they [the students] have. Some of the students are very bright and some of the students are very intelligent, that is there. But they have the difficulty to [linguistically] adapt to new subjects.

However, two teachers from the IT and IBA departments maintained that the English language levels of the students were generally irrelevant to their academic achievement. Interestingly, both of the teachers were Omanis. They stated their views as follows.

> **Teacher 18**: I do not think that there is a direct relation between speaking well and getting good scores [in IBA courses]. I have students who cannot communicate with me in English but they get good scores in the final exam ... But this is not true in all courses, for example in the accounting course, students deal only with numbers; language has nothing to do with

it. But, in this course *Business Fundamentals* the students have to write two assignments, so it [proficiency in English] might affect their marks.

Teacher 17: Most of students who get high scores in my course, their English is not that good. And most of the students who come into the college with good scores in high school certificates their English language is not that good and then it improves in the college. It depends on the students themselves if they want to study or not.

9.3.2. Students' Linguistic Readiness for the First Year Courses

According to most of the FY teachers, their students seemed to be linguistically unready for the FY academic courses. This view was derived from their classroom observation of the students' linguistic inability to accomplish certain tasks. Both English language and academic courses teachers shared some of the problems students face in FY including: poor communication skills, inability to digest lectures, inability to read assigned materials and other poor study skills such as summarizing and note-taking.

Teacher 14: I just teach writing, I do not have all of the other skills. It is difficult because they [students] are quite of low level. The books in this level are too difficult for them.

Teacher16: They are not ready. It is not only about the language, it is also about the study skills. They do not have study skills at all.

Teacher 18: According to my course, *Fundamentals in Business*, and [according to the fact that] they are First Year students and first year in the Business specialization, I think [that] they are very weak in the English. Because I am Omani and they always say to me 'please, explain it in Arabic'.

Teacher 13: The book that is designated to the course by the college, only 5% of the students can read it. Because of that I had to go throughout the book and write it in my own notes which are more appropriate to the average level of the students.

Teacher22: Honestly speaking 80 to 90% of the students are not up to the level of the first year. This is my opinion and this is my colleagues' opinion too when we speak about it. Students mostly cannot express themselves, we always communicate in sign language. They cannot pronounce simple, simple words.

As there is not sufficient space to include all of the teachers' comments on the difficulties the students seemed to encounter in FY academic courses, the table below summarises and categorises them into study skills and language skills.

Table 9.3 Study Skills and Language Skills Difficulties Faced by FY Students

Study Skills	**Language Skills**
Reading effectively and making summaries	Grammar and spelling
Listening for gist and taking notes	Structuring sentences
Paraphrasing answers	Pronouncing specialised vocabulary
Accomplishing assigned homework	Reading designated texts
Attending class on time	Fluency in speaking
Participating in class discussions	Knowing basic specialised vocabulary
Presenting efficiently	Writing in Year One level

In spite of the above statements that indicated most students' inability to cope with the FY courses, three teachers perceived the difficulties students faced as being part of the normal educational path and felt that they should not be of concern. They seemed to think that the students were quite capable of meeting the requirements of the First Year courses regardless of the challenges. All of these teachers were from the English language department.

> **Teacher 1:** I can speak from my experience both within the classroom and outside the class room. The learning outcomes for the students or better yet for the course … the students meet that profile, those criteria. Students are able to fully express themselves without hesitation whatsoever, generally speaking.

> **Teacher 6**: There will always be those who are not ready for the First Year even when they are in [the] second year, still not ready for [the] first year. There are slips, slides, [and] hurdles to cross; there are some for people to get over.

> **Teacher 13:** If I think of my 30 students in the first year, I have 5 students who have really low level of English and they still earn roughly 60% in my quizzes and exams. Which means yes they are still able to understand the content, even with limited English.

9.3.3. Teachers' Perceptions of the Effectiveness of the Foundation Programme

When the teachers were asked about their perceptions of the students' readiness for FY courses linguistically, almost every one of them answered with negative comments and claimed that the FP was ineffective in equipping the students with the

language skills required for the FY study. Some of them seemed to believe that this ineffectiveness sometimes was caused by inappropriate and very lenient assessment criteria. In the following extracts, the first two teachers mainly criticised FP curriculum, and the following three teachers argued that the assessment was not strict enough.

> **Teacher 10:** Actually I do not know anything about the Foundation Programme or how they [students] have been taught. But I can tell you about the standard of the students. They are very good in terms of how they think but their English language is very weak, it stops them from giving out ideas, making up ideas or sharing their ideas.
>
> **Teacher 16:** I do not know … I can tell you that there is a gap between the two years and there is a problem. Students pass when they do not deserve to pass. Most of them do not have appropriate knowledge of the grammatical rules.
>
> **Teacher 22:** I would say where it goes wrong is in the foundation, if they were given a proper Foundation Programme we would not have all of the problems that we are talking about now… the Foundation Programme should be stricter.
>
> **Teacher 12:** they [students] go from the foundation year to the first year and this is something I have a bit of issue with, having to get only 50% of the total mark, 50% is a low percentage to get to another level. This is will be considered a fail in another context and even 60% this means that they have the basic skills in order to understand English, so if you put the same students with a business book which is probably 400 pages then it becomes much more difficult.

In the interviews, five teachers mentioned that the female students showed a better command of the English language than the male students and argued that they consequently were more able to meet the linguistic requirements of FY courses and excel academically.

> **Teacher 17:** Also the girls are always better than the boys in terms of English language, I do not know why but it is the case here. I do not think that it depends on the first year or Foundation Programme teachers; I think that it depends on the fact that they [male students] don't take the courses seriously. They think of the Foundation Programme as a pass or fail so they do not have to study hard, it will affect them in the First Year.
>
> **Teacher 9:** Girls tend to really study hard and a lot. So most of the time, they are really, really good and much better than the boys.

Also, eight teachers seemed to believe that what they considered negative attitudes and low motivation towards learning could be impeding students' ability to perform well in the FY assessment.

> **Teacher 10:** It is not about English, it is also about the attitudes and the way you take it anywhere. Some students are into the negative aspect of westernism and that language and westernism they correlate and that's one reason why they do not like the language

> **Teacher 20:** [Students] are not prepared for the college environment, they are just not used to its pressure and they do not have a very good standard of English when they come to the college. So it is difficult for the Foundation Teachers to deal with them, I speak from experience.

9.3.4. Assessing Language Accuracy of Written Assignments in Academic and English Language Courses

When asked about how the language of written assignments was evaluated, the teachers expressed mixed views. Almost all teachers of academic courses maintained that it was necessary to ignore the language weakness of written assignments because of the students' low language levels. They appeared to believe that marking the language of written assignments would lead to giving the students very low marks. On the other hand, the English language teachers' views seemed to be split, some indicating that the content of a written assignment should be the main focus of assessing written assignments, others maintaining that their job was to mark the language of written assignments. Very few English language teachers stated that marking a written piece should consider both content and language using a making scale. These views are presented below.

9.3.4.1.Focus on Language

> **Teacher 20:** For me because I am teaching them writing, I am concentrating more on the grammar and structure of the language rather than the content.

> **Teacher 12:** I cannot say anything about IT knowledge, I cannot say if this student know or does not know things in IT but I can evaluate in the language side of it.

9.3.4.2.Focus on Content

Teacher 10: I do not focus on grammar. No. I cannot afford to focus [on grammar]... There are the ideas [in the written piece] and you mark it based on its ideas and meaning.

Teacher 18: If I do and according to their grammar and spelling, they will all get zeros. So in my course I always score them on the information or what they understood, and the content is more important to me.

Teacher 17: I do not consider grammar or spelling or any language related aspect because If can understand what they want to say, I do not mark them for grammar and they always ask me whether I mark them for grammar and I say no.

Teacher 1: No it is more of content, it is more of the students actually if the students [are] actually able to convey meaning and I am actually able to digest what the students are trying to address.

9.3.4.3.Focus on Content and Language

Teacher 5: In the test you have a marking rubric and they put a lot of focus on content and mechanics, spelling…so most of them will pass if you follow the rubric ... So by the time you finish awarding scores for each element, it is a pass mark. And then I could say that this person has got 12 out of 20 and you could say my goodness I would have given him 7 or 5 but if you follow the rubric, this is what you get.

9.3.5. Problematic Issues in Marking Written Assignments

Throughout the interviews, problematic issues were raised about marking written assignments, not only in the academic departments, but in the English language departments too. In all academic departments, most of the teachers seemed aware of the lack of clear criteria for marking language accuracy in written assignments. It was also maintained that sometimes there was no consensus between different teachers of a course on how language accuracy should be considered in marking written assignments.

Teacher 3: No there is not a common scale or way that we use. Some of the teachers are very strict, they check every single letter and they are always taking scores off for every spelling mistake. They are very strict.

Teacher 2: I do not know if the other teachers are doing the same or not because every teacher is doing their marking by themselves, so no … there is no … I do not know what they are doing.

One other issue that was noticed in the English language department was that most teachers did not mention the role of the marking scale provided by the ministry to evaluate written assignments. The very few who mentioned it described it as lenient or unsuitable for evaluating scripts. When asked about how the students' essays were assessed, the following answer was given. These issues could offer some explanation of the ambiguity and uncertainty on the students' side about the role of the language skills in assessing written assignments expressed in Section 9.2.4 and reflected in the responses to the student questionnaire in Chapter 8.

9.4. Discussion

This section compares the results obtained from the focus groups and interviews and discusses the main shared issues. There were four main topics that both the students and their teachers raised: the correlation between language proficiency and academic achievement, language related difficulties students face in FY, effectiveness of FP assessment in retrospect, and assessing language accuracy in written assignments of FY English language and academic courses.

9.4.1. Correlation between Language Proficiency and Academic Achievement

In the student focus groups and teacher interviews, most speakers seemed to believe that performance in the English language assessment predicts achievement in academic courses. The justifications that the students and teachers offered for this belief were similar. However, a few students and teachers maintained that proficiency in English language was not the major determinant of achievement in academic courses; these participants were mainly from Rustaq College, specifically several of the IT and IBA teachers and students. This finding is in accordance with the findings of previous studies on predictive validity that have considered the students' fields of study, and which have found that the language requirements of the different areas of studies varied, and consequently the academic difficulty

experienced by the students and the appropriate language level envisaged by the teachers varied as well (e.g., Bayliss, 2006; Lynch, 2000).

The participants from Sur College seemed to believe that proficiency in English language was a key factor influencing achievement in the academic courses. This finding is not surprising and accords with evidence from other studies that indicate that proficient users of a second language tend to perform better in linguistically demanding courses (Woodrow, 2006; Powers , Kim, & Weng, 2008). The results of the teacher and student questionnaires indicated similar conclusions and showed a strong agreement with the views that proposed a positive relationship between language proficiency and academic achievement. The tendency of Rustaq teachers to consider English language proficiency as less important than the other teachers could be understood by looking at the different groups that constituted the Rustaq sample: most of the teachers were from the IBA department. As has been pointed out in Chapter 8 and will be raised again in Chapter 10, IBA teachers and students considered English language proficiency as one of many factors that contributed to academic achievement and did not see it as the major one as the other teachers and students did. Similarly, the predictive validity of FP assessment for the IBA student group was lower than that of the CS, English or IT groups.

9.4.2. Language Related Difficulties the Students Face in FY

Furthermore, most of the students and teachers agreed that the students were not ready for the FY courses linguistically. They mentioned a number of language related difficulties such as: reading courses' textbooks and writing extended essays. The teachers' comments included these two difficulties but added several more. They felt that the students were weak not only in receptive and productive language skills, but also in vital study skills required in FY. Similar findings were reported by Hill, Storch and Lynch (1999), who maintained that the difficulties faced by the students in their academic studies are not limited to language related factors but also include what they called, "study related factors" which included difficulties with time management and workload. They claimed that these non-linguistic factors could

explain the low IELTS predictive power of academic achievement. Furthermore, Xu (1991) found that English language proficiency was the best predictor of students' ability to cope with the difficulties of academic study; and reported that the perceived difficulty of academic study positively correlated with the students' self-evaluation of language proficiency. Similar studies on other factors that affected academic achievement are presented in Chapter 3 (see Section 3.4). These findings will be revisited in Chapter 10 when discussing the findings on the predictive validity of FP assessment.

9.4.3. The Effectiveness of FP Assessment in Retrospect

Another point that was raised by both the students and teachers was the perceived leniency of FP assessment criteria. Most of the students mentioned that they had thought that FP study would be more intensive and that FP assessment and criteria would be stricter. Several of them felt that the quality of FP should be reconsidered to include more EAP activities and generally indicated that the language requirements for CAS admission should be stricter. These perceptions actually conform to the ones generated by the questionnaires (See Section 8.2.3).

9.4.4. Assessing Language Accuracy in Written Assignments

When the students and teachers were asked about whether the language accuracy of academic and non-academic written assignments was/should be assessed, the responses varied. Most of the teachers of the academic courses stated that the language mistakes were usually overlooked when marking written assignments. They also mentioned that there was no common scale or policy on language marking in the academic courses, though; they sometimes discussed how language will be marked with their peers. This view accords with the teachers' responses to a questionnaire item that entailed their tendency to overlook language accuracy when the content was comprehensible (see Section 8.4).

Most of the students stated that whether language accuracy was used as a criterion in marking written assignments was not clear to them, however, most of them guessed that language was actually considered in marking academic assignments. Few of the students seemed to feel that language was not and should not be a criterion in marking written assignments in the academic courses. But if it was, they conceded

that they would have had paid more attention to the language accuracy of their written assignments. Their views were in accordance with the findings obtained from the student questionnaire on students' uncertainty about marking criteria (see Section 8.2.3). They also mirrored the results of Norton and Starfield (1997) who reported that the students at Wits University in South Africa seemed uncertain about how much language was considered as a criterion in assessing the academic compositions; and that this uncertainty was equally shared by their teachers.

The English language teachers' responses were split three ways about considering language and content in marking written assignments; some regarding language and content to be of similar importance, others considering language or content to be more important than the other. Interestingly very few of the English language teachers referred in their responses to the marking scale that was supposed to be used for marking students' assignments. Similarly, Hay and MacDonald (2008) reported that the physical education teachers in their study relied on own memorised representations of the marking criteria which the authors argued compromised the validity of the assessed construct and inter-rater reliability. This concern was similarly raised by the Phase 1 teachers who seemed to believe that not all of the teachers used the same criteria in marking and when they did, they did not use it systematically (see Section 7.3.4.1).

Nonetheless, those few teachers who referred to the marking scale in their responses were critical of its validity. They explained that it overlooked issues such as plagiarism and actual students' language levels. They mentioned that its criteria were surprisingly easy to meet and therefore, the students were able to pass effortlessly from one level to the next. This conclusion was similarly reached by analysing the writing marking scale in (see Section 5.3.2.2). Previous studies on language assessment predictive validity that included formative assessment recommended a "rigorous process of cross marking" to increase the reliability of the assessment (Cope, 2011). The students' responses to the same topic made some references to the criteria of the marking scales used. Interestingly, they believed that avoiding plagiarism and abiding by the stated word limit were the most important criteria in

the marking scale. A few students, however, were not sure about what was included in the marking scales.

9.5. Summary and Concluding Remarks

In this chapter, the findings obtained from student focus groups and teacher interviews were categorised into four common themes: correlation between language proficiency and academic achievement, language related difficulties the students face in FY, effectiveness of the FP assessment in retrospect, and assessing language accuracy in written assignments.

Some of the findings presented in this chapter are similar to the ones obtained via different methods (i.e., document analysis or questionnaires) and presented in previous chapters. One of them is the apparent difference in opinion about the importance of proficiency in English in academic achievement according to participants' specializations. Participants from the IT and IBA departments agreed that there is a positive relationship between proficiency in English language and academic achievement but to a lesser degree compared to the participants from the CS and English language department. These views support the findings on predictive validity of the FP assessment as presented in Chapter 11.

This chapter revealed that, according to the participants, the FP content and assessment criteria should be changed. The FP content was considered to be very easy by the most of the students who tended to express that a more challenging programme would be better for improving their English. They also compared the FP in CAS to other programmes in other higher education institutions maintaining that their programmes were better because they were more condensed and challenging. The teachers implied a similar perception and most of them identified the difficulties students faced in the FY study to be related to both language skills and study skills. This means that when changing the FP content and exit criteria to better suit the needs of the FY study, study skills should be as well considered and included in FP.

The findings also revealed that there were unclear or non-existent guidelines on how to consider language accuracy of written assignments. Most academic courses teachers indicated that they ignored language when meaning was comprehensible,

and English language teachers expressed different views on this matter. Most of the students seemed to be unclear about the criteria but maintained that they would have paid more attention to reviewing the language of their assignments if they had known that it was considered in marking. The implications of these findings are discussed in Chapter 11.

Chapter10: Predictive Validity of the Foundation Programme English Language Assessment

10.1. Introduction

This chapter attempts to assess the predictive validity of the Foundation Programme (FP) English language assessment; more specifically how well students' scores in the FP assessment predict their scores in First Year (FY) academic courses (i.e., Information Technology (IT), International Business Administration (IBA), and Communication Studies (CS). It also investigates the predictive validity of the assessment of each the General English Skills (GES) and Academic English Skills (AES) separately; and identifies differences in the predictive validity of the FP assessment across the groupings by gender, college, specialisation and self-evaluation. The questions that this Chapter addresses are presented in Box 10.1.

Box 10.1. Study Questions Addressed in Chapter 10

3. What was the predictive validity of the English language assessment for student performance on the academic courses?
3.1. Did student performance in English language assessment in the FP correlate positively with the performance in academic courses?
3.2. Did the strength of correlation between the language proficiency and academic achievement differ significantly when students' scores in English language tests only, or continuous assessment only, were used, instead of the overall scores in both?
3.3. Did the groupings by college, gender, self-evaluation and specialisations show significant differences among the correlations between language proficiency and academic achievement?
3.4. How much did the teaching and assessment in the First Year academic courses depend on students' language proficiency?

The first section of the chapter starts with a presentation of the operational definitions of "language proficiency" and "academic achievement". Next, it explores the predictive validity of FP English language assessment, first for the whole sample, and then for specific groups. After that, it investigates the predictive validity of assessment in the last year of high school and compares it with the predictive validity of FP assessment. This is followed by an evaluation of the appropriateness of the current cut-off point in the light of the findings on the predictive validity across the groupings by specialisation. The last section briefly recaps on the language demands of the IT, CS and IBA introductory courses in the FY as discussed in Chapter 5 and explains the variations in the FP predictive validity amongst the student specialisation groups, arguing that certain specialisations are more linguistically demanding than others.

10.2. Operational Definitions of 'Proficiency' and 'Achievement'

Before investigating the relationship between the students' language proficiency and their academic achievement, it is crucial to explain how the concepts 'language proficiency' and 'academic achievement' were operationalised. Students' English language proficiency was represented by their average grades on the two FP English language courses (i.e., AES and GES). Likewise, the students' achievement in academic courses was represented by their average grades in the FY academic courses in the first semester. Scores on courses unrelated to the specialisations or on those taught in Arabic (e.g., Islamic Culture or Omani Economic History) were not included in calculating the students' average scores on the academic courses.

Another point to clarify is how the Grade Point Average (GPA), used in CAS to report students' achievement, was employed in this study. The GPA is "the Grade Point Average of the numeric value of the entire results that the student has passed or failed in that semester" (CAS, 2010, p. 4). To calculate the GPA, students' scores were transformed from numeric grades to grade points ranging from 0 to 4 using the scale in Table 10.1, which is also the standard scale for calculating GPA in CAS. The crude GPA form of the FY was deemed to be unsuitable for this study as it included the average results of all courses taken in a specific semester; this study was looking

only at English language medium courses that were related in content to the students' academic specialisations. Therefore, only the grade points of the academic courses that were taught in English and related in content to the students' academic study were included in the GPA used to represent academic achievement.

One complication encountered was that the students' scores in the academic courses were only available in a grade point system, while their scores in the FP assessment were available in a numeric system. To overcome having the grades in two forms, scores in the FP were converted to grade points using the scale used in CAS and shown in Table 10.1. For example, if a student's score in the FP was between 80 and 84, this score was converted to a grade point of 3.0.

Table10.1. Conversion Table for Scores Used in CAS*

Numeric Grade	**<50**	**50-54**	**55-59**	**60-64**	**65-69**	**70-74**	**75-79**	**80-84**	**85-89**	**90-94**	**95-100**
Grade point	0	1.0	1.3	1.7	2.0	2.3	2.7	3.0	3.3	3.7	4.0
LetterGrade	F	D	D+	C-	C	C+	B-	B	B+	A-	A

*. from the Registration Office at Sur CAS, personal communication, February 14, 2012

10.3. Predictive Validity of FP Assessment

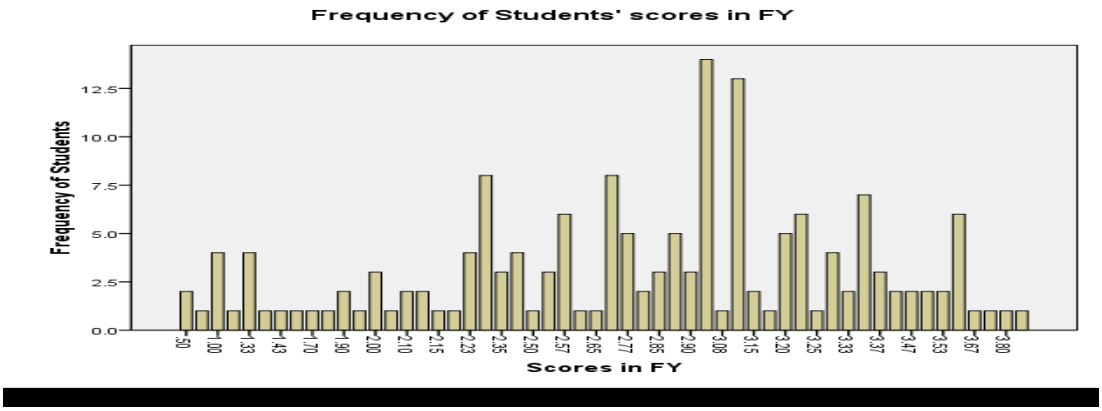
The predictive validity of a test is a measure of how well it predicts some future performance, though not always in another test or tests. In this study, students' grades in the FP assessment in two CAS colleges were correlated with their grades in the academic courses of the first semester of their FY, which actually was in the following academic semester; students started the FP in February 2011 and the FY in September 2011.

The sample started out with 184 students on the FP, and then it decreased to 176 students in the FY due to several factors discussed in Chapter 8. In this chapter, the size of the sample included in the statistical tests to investigate the correlations decreased further to N=163 because, firstly, not all students' took specialised academic courses in the first semester of the FY as some of them studied general or Arabic medium courses only; secondly, not all students' scores in the academic courses could be obtained as some of the students withdrew from certain courses. Therefore, the sample in the subsequent sections includes only 163 CAS students.

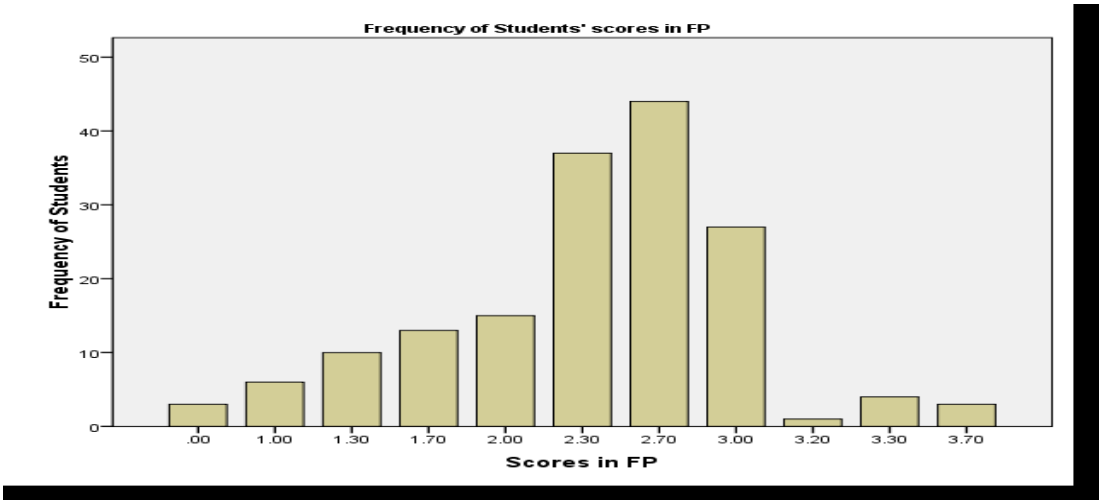
10.3.1. FP Predictive Validity for the Whole Sample

Students' grades in the FP English language courses and their average grades in the FY academic courses were tested for normality of distribution using Kolmogorov-Smirnov, Shapiro-Wilk tests, and histograms. The results showed that the students' scores were negatively skewed (see appendices 10.1 and 10.2). For this reason, only non-parametric statistical tests were used to explore the data set. The following graphs show the distribution of students' grades in FY and FP.

Graph 10.1. Distribution of Students' Scores in FY



Graph 10.2. Distribution of Students' Scores in FP



A Spearman's rank correlation was used to explore the predictive validity of the FP by correlating the students' grades in the FP assessment and their grades in the academic courses assessment. The results showed a highly significant but only

moderately strong correlation between the two variables, rho=0.31, p < 0.01 (see Table 10.2.). In addition, the difference in the predictive validity of each of the FP courses (i.e., GES and AES) was explored. The students' grades in the GES assessment correlated with their average grades in the academic courses moderately, rho =0.37, p < 0.01. However, the correlation between the students' grades in the AES assessment and in the academic courses assessment was weaker, rho=0.27, p < 0.01. In other words, the students' grades in the FP courses are generally a moderate predictor of their grades in the academic courses, but better predicted by their grades in the GES assessment. It is worth remembering at this point that the GES assessment consisted of standardised tests while the AES assessment consisted of performance assessment tasks.

10.3.2. Comparing the Predictive Validity of the FP across the Groups

10.3.2.1. Differences between College Groups

The predictive validity of English language assessment in the FP was stronger for the participants from Sur College than it was for those from Rustaq College. The table below shows that Spearman coefficients for the students' grades in the FP and FY assessment were rho= 0.46, p = 0.002 for Sur College (*N*=44); and rho= 0.16, p=.088 for Rustaq College (*N*=199). It is worth noting that the correlation between the scores in the FP and FY assessment was found to be non-significant in Rustaq College (see Table 10.5).

Table 10.3. Correlation between Scores in FP and FY Assessment by Colleges

College	Correlation	Sig.	*N=163*
Rustaq	.16	.088	199
Sur	.46**	.002	44

**. Correlation is significant at the 0.01 level (2-tailed).

10.3.2.2. Differences between Gender Groups

The correlations between the students' scores in the FP assessment and their grades in the FY academic courses assessment were not very different between the gender groups. The Spearman coefficient for the male group was rho = 0.30 and for the female group rho = 0.32, at significance levels of p =0.07 and p < 0.01 respectively.

Table 10.4. Correlation between Scores in FP and FY assessment by Gender

Gender	Correlation	Sig.	*N=163*
Male	.30*	.07	61
Female	.32**	.000	101

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

10.3.2.3 .Differences among Self-evaluation Groups

As explained in Chapters 6 and 8, the students had been asked to self-evaluate their language proficiency using the descriptors: weak, average, good, very good, and excellent. As only one student evaluated his language proficiency as weak, this group was added to the average group to be able to conduct statistical tests. The Spearman correlation between students grades in the FP assessment and their grades in FY academic courses assessment ranged from rho=0.17 in the *average* group to rho=0.88 in the *excellent* Group (see Table 10.5). This means that the higher the students evaluated their language proficiency the stronger the predictive validity coefficient became, and consequently the more their performance in the academic courses became predictable by their performance in the FP courses.

Table 10.5. Correlations between scores in FP and FY assessment by to Self-Evaluation Groups

Self-Evaluation	Correlation	Sig.	*N*=163
Average	.17	.59	12
Good	.25*	.02	85
V. Good	.39**	.005	51
Excellent	.88**	.009	7

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

10.3.2.4. Differences among Specialisation Groups

Interestingly, the strength of predictive validity of the FP assessment varied depending on the students' specialisations. Table 10.7 shows that the students' grades in IBA and IT courses were less well predicted by their grades in the FP assessment than were their grades in CS and English language (education) courses. The predictive validity of FP assessment in the specialisation groups ranged from rho = 0.18, p = 0.12 for the IBA group to rho = 0.64, p<0.01 for the CS group (see Table 10.6).

Table 10.6. Correlations between Scores in the FP and FY Assessment by Specialisations

Specialisation	Correlation	Sig.	*N=163*
Information Technology (IT)	.41*	.008	41
Communication Studies (CS)	.64**	.002	21
International Business Administration (IBA)	.18	.12	78
English Language-Education	.57**	.005	23

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

The difference in the predictive validity between the two Colleges could be explained by the type of specialisations taught in each of the colleges and the size of student samples represented by each specialisation in this study (see Table 10.7, Figures 10.1. & 10.2). The participants from Sur College specialised in either IT or CS; and the participants from Rustaq College specialised in IT, IBA or English language (Education major). The fact that most of the Rustaq College participants were IBA students (66.93% of the sample), and that the predictive validity of FP asessment for the IBA group was non-significant, could very well explain the non-significant result obtained for the predictive validity of the FP assessment in this group.

Table 10.7. The FP assessment Predictive Validity by College and Specialization

College	**Specialisation**	**Correlation**	**Sig.**	***n***
Rustaq	IT	.27	.27	18
	IBA	.11	.31	78
	English Language-Education	.66**	.001	23
Sur	IT	.14	.52	24
	CS	.73**	.000	21

**. Correlation is significant at the 0.01 level (2-tailed).

Figure10.3. Student Distribution by Specialisations in Sur College

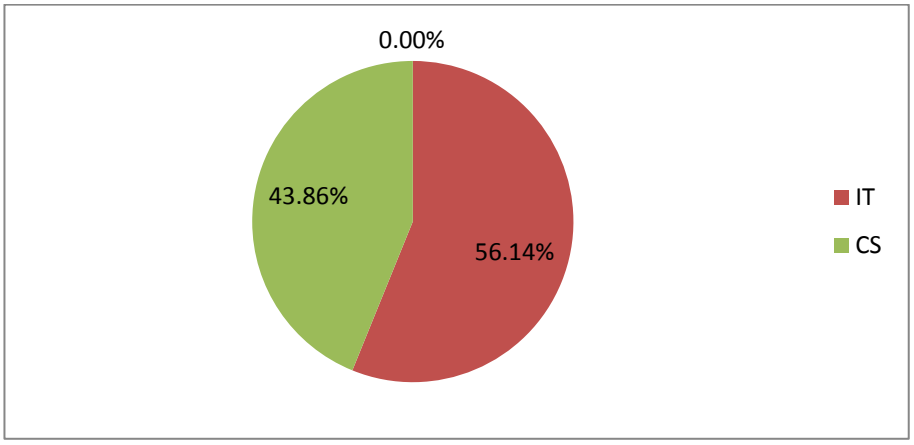
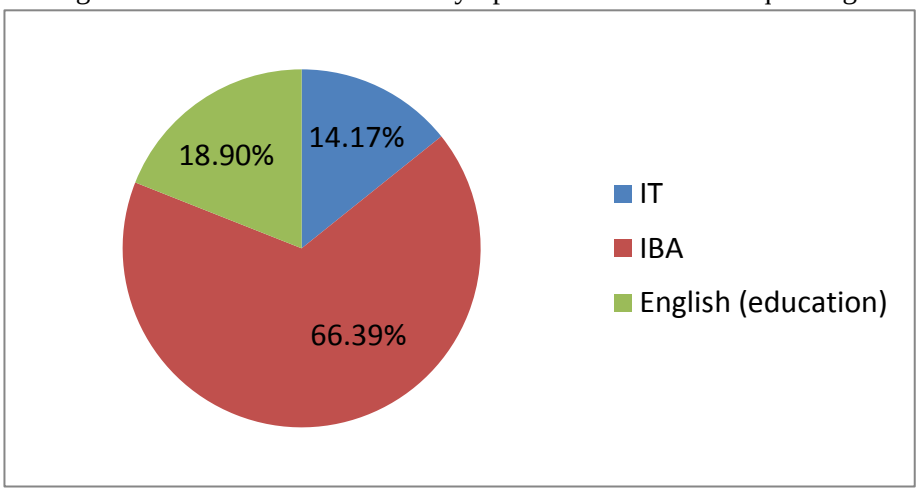


Figure10.4. Student Distribution by Specialisations in Rustaq College



10.3.3. Academic Achievement as Predicted by Students' Scores in High School

Students' average grades in the last year of high school (in both academic courses which are taught in Arabic, and English language courses) were correlated with their grades in the FY academic courses to investigate if the former were better predictors of the academic achievement in the FY than were the FP grades. The correlation between the students' average grades in high school assessment and their average grades in the FY assessment (rho= 0.37, $p < 0.01$) was, in fact, similar to the correlation between their grades in the FP assessment and their grades in the FY assessment (rho= 0. 311, $p < 0.01$).

The predictive validity of high school scores with regards to academic achievement was comparatively strong for the IT and CS groups (see Table 10.8). However, the

predictive validity of high school scores for the IBA group and English language (Education) group was non-significant.

Table 10.8. Correlations between Grades in High School Assessment and Grades in FY Assessment by Specialisation

Specialisation	Correlation Coefficient	Sig.	*n*
Information Technology	.56**	.000	41
Communication Studies	.62**	.003	21
International Business Administration	.12	.21	77
English Language-Education	.31	.27	21

**. Correlation is significant at the 0.01 level (2-tailed).

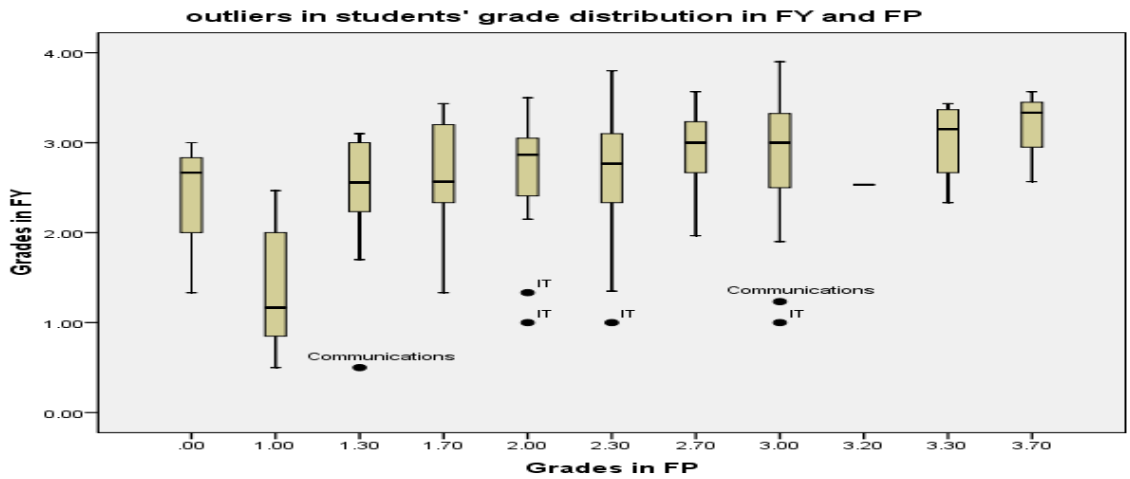
10.3.4. FP Cut-off Point and Academic Achievement

In the light of the above findings, this section discusses the appropriateness of using one cut-off point for the four specialisations in the FP assessment. The appropriateness of the FP cut-off point was explored by cross-tabulating the students' grades in the FP assessment and their grades in the FY assessment; the students were divided into four groups based on their specialisations.

It was essential to decide before starting this process what should be the minimum acceptable grade in the academic courses. Determining such a grade was hoped to facilitate identifying the FP grade cut-off point that could most likely predict success in the different specialisations and lessen false positive cases (i.e., students whose FP grade is higher than the cut-off point but who fail or struggle in their FY academic studies). As the CAS Academic Regulations state that "a student who achieves a semester Grade Point Average of less than 2.00 will be placed on probation in the following semester" (p.20), the grade point 2.00 seemed to be a good indicator of acceptable academic achievement (i.e., a threshold of academic success). Therefore, the grades of students' on academic courses that were below or above the grade 2.00 were scrutinised for the most common corresponding grade point in the FP assessment. Also, the students' grades in the academic courses were scanned for those who scored 1.00 (i.e., established cut-off point in FP assessment) to track their achievement through their grades in the academic courses.

Table 10.9 shows the distribution of the students' grades in the FP and their corresponding grades in the academic courses. These figures suggest that the current cut-off point, which is 50% of the total FP score (i.e., an equivalent score to 1.00 in the grade point system), is appropriate for the IT, IBA and English language groups where most of the students with the minimum entry grade of 1.00 passed and sometimes performed satisfactorily by obtaining a grade point of 2.00 or higher in the academic courses assessment. There were few cases in which the students obtained a grade point of 1.00 in FP assessment but received a grade point of less than 2.00 in the academic courses assessment. In IT, there were five such cases and, in IBA, there was one such case, but none were found in English language (education). The following Graph shows the outliers in FP and FY scores by specialisation.

Graph 10.5. Box-Plots of Student Scores Distribution in FY and FP by Specialisation



However, generally CS students who scored 1.00 or less in the FP seemed to struggle in CS academic courses and mostly managed to obtain the pass grade only (i.e.,1.00). Almost all of the other CS students who scored (1.3) or higher in FP were able to score (2.00) or higher in the CS academic courses. Given the strong predictive validity of FP assessment when it comes to achievement in CS courses (rho = 0.62, $p < 0.01$, n=21), and the tendency of the students who scored 1.3 or higher in FP to

obtain 2.00 or higher in the CS courses, it is tentatively suggested, subject to investigation of a larger sample and appropriate consultation, that the entry level for the CS should be changed to 1.3 (i.e., 55-59% of FP assessment total score).

Table 10.9. Distribution of Students Grades in FP assessment and Academic courses Assessment by Specialisation Groups

Specialisation			**Average Grades in Academic Courses**										
			1.0	1.3	1.7	2.0	2.3	2.7	3.0	3.3	3.7	4.0	**Total**
IT	Grade in FP	1.00	0		0	0	0	1	0	0	0	0	1
		1.30	0		1	0	1	0	0	0	0	0	2
		1.70	0		0	1	0	0	0	2	0	0	3
		2.00	1		0	0	0	1	1	0	1	0	4
		2.30	1		1	1	0	2	0	2	1	1	9
		2.70	0		0	0	0	0	2	3	1	0	6
		3.00	1		0	0	1	0	2	1	0	1	6
		3.30	0		0	0	0	0	0	0	1	0	1
		3.70	0		0	0	0	0	0	0	1	0	1
	Total		**3**		**2**	**2**	**2**	**4**	**5**	**8**	**5**	**2**	**33**
CS	Grade in FP	1.00	3	0		0	0	0	0	0			3
		1.30	1	0		0	1	0	0	0			2
		2.30	0	0		0	1	1	2	1			5
		2.70	0	0		1	0	1	3	0			5
		3.00	0	1		0	0	0	0	1			2
	Total		**4**	**1**		**1**	**2**	**2**	**5**	**2**			**17**
IBA	Grade in FP	.00			0	0	0	0	1	0	0		1
		1.30			0	0	0	0	0	1	0		1
		1.70			0	0	0	1	0	0	0		1
		2.00			0	0	1	1	3	0	0		5
		2.30			1	1	2	2	3	2	2		13
		2.70			0	0	3	6	5	8	2		24
		3.00			0	1	0	2	2	3	2		10
		3.20			0	0	0	1	0	0	0		1
		3.30			0	0	0	0	0	1	0		1
		3.70			0	0	0	1	0	0	0		1
	Total				**1**	**2**	**6**	**14**	**14**	**15**	**6**		**58**
English Language (Education)	Grade in FP	.00				0		1	0	0	0		1
		1.00				1		0	0	0	0		1
		1.30				0		0	1	1	0		2
		1.70				0		1	0	2	1		4
		2.00				0		0	1	3	0		4
		2.30				0		1	0	0	2		3
		2.70				0		0	0	1	0		1
		3.00				0		0	0	0	2		2
		3.30				0		0	1	0	0		1
	Total					**1**		**3**	**3**	**7**	**5**		**19**

There are two main implications for the suggested increase of entry level for the CS students. First, CS programme might become less appealing to students because of its linguistically demanding nature in addition to the fact that CS graduates are less employable compared to IT, IBA, and Design graduates (see Section 1.4.1). Second, increasing the cut-off point might not be welcome by policy makers as it means spending more money and providing extra resources for students at FP to be able to meet the suggested higher level of English language proficiency. Students at CAS receive full scholarships, so CAS will be financially liable for any extended time spent learning English language.

10.4. Language Demands of Different Specialisations

As has been discussed in Chapter 5, the language demands of studying and being assessed in the FY academic courses differs from one specialisation to another. An analysis of the course learning outcomes, assessment specifications and test papers (of Spring 2009) revealed that the CS learning outcomes and assessment instruments require more command of English than do the learning outcomes and assessment instruments of IBA or IT (see Section 5.3.5). This finding conforms to the findings on predictive validity that indicated that scores in the FP assessment correlated more strongly with the scores in CS courses than they did with the scores in the IT or IBA courses.

10.5. Discussion

The results presented in this section are discussed in the context of relevant literature. It is divided into three main subsections; the first deliberates on the general findings on FP predictive validity, while the second and third discuss the differences in predictive validity across specialisation groups and self-evaluation groups.

10.5.1. Predictive Validity of FP

Investigating the predictive validity of the English language assessment in the FP showed a significant but low correlation between the students' grades in FP English language courses and their academic courses in the FY rho= 0.31. GES grades showed a stronger correlation coefficient (rho=0.36) with the academic courses grades than did AES grades (rho=0.27). This finding is in line with the conclusions drawn from similar previous studies conducted on the predictive validity of various English language tests that were used as gatekeepers to higher education institutions

such as IELTS, TEAM, and various local tests (Davies, 1990; Elder, 1993; Cope, 2011; Lynch, 2000). Though these studies varied in the sample sizes, students' specialisations, levels of higher education, and measures of language proficiency and academic achievement, most of them concluded that the correlation between English language proficiency and academic achievement was weak to moderate, between 0.2 and 0.4.

10.5.2. Predictive Validity of FP across Specialisations

This study found that the strength of the correlation between the students' language proficiency and academic achievement varied considerably depending on the students' specialisations. These different predictive validity values for the specialisations could be partly explained by the language demands of these courses as reflected in their stated learning outcomes, assessment instruments and test tasks, as discussed (see Section 5.4). The CS assessment instruments and test tasks seemed to draw upon students' language skills more than did those of the IT or IBA assessment instruments.

Furthermore, the variance in the strength of the predictive validity of language assessment across specialisations was similar to that reported in several previous studies. Jochems et al. (1996) found that the value of the predictive validity varied from r= 0.32 to r= 0.46 in Computer Sciences and Engineering majors. Their study looked at the correlations between Dutch language proficiency as a second langauge (it was the medium of study) and academic achievement. In the English language domain, Cope (2011) reported that the value of the correlation varied between different disciplines when he studied the predictive validity of three types of English language entry programmes. Lynch (2000) found that there was some difference in the correlation coefficient between the English language test used at the University of Edinburgh and students' average scores in the academic courses across the students' different fields of study. For example, the correlation coefficients in the Arts and Veterinary Medicine were non-significant, whereas, the coeffcients in Social Sciences, Law, Science and Engineering were r= 0.23, r= 0.32 and r= 0.24 respectively. Similarly, Huong (2001) claimed that the correlation between language proficiency and academic achievement in the linguistically demanding disciplines

(e.g., TESOL) was stronger than it was in the less linguistically demanding disciplines (e.g., Engineering). Woodrow (2006) reported the correlation coefficient between the students' bands in IELTS and their GPA in TESOL courses to be r= 0.4, p<0.01, *n*= 62. Other similar findings on the predictive valdity of langauge assessement in higher education contexts were reported in Section 3.4.

10.5.3. Predictive Validity of the FP across Self-evaluation Groups

The correlations between language proficiency and academic achievement seemed to differ according to the students' self-evaluations of their language abilities. The higher the students evaluated themseleves, the stronger the correlation between their grades in FP assessment and academic courses assessment became. Very few studies on predictive validity have investigated the possible contribution of the students' self-evaluations to the strength of the predictive validity of presessional language assessment (Powers, Kim, & Weng, 2008; Xu, 1991), the second of these produced interesting results. Xu (1991) investigated the correlation between students' self-evaluations of their language proficiency and self-reported academic difficulties, and the correlation between TOEFL scores and self-reported academic difficulties. Xu found that the students' self-evaluations were a better predictor of the percieved academic difficulties than were TOEFL scores. Though Xu's focus was on perceived academic difficulties, his findings draw attention to the role of self-evaluation in understanding possible future academic difficulties. Given these previous findings and considering the findings of the current study, it is suggested that self-evaluation should be considered a possible variable in English language assessment predictive validity in future research.

10.6. Summary and Concluding Remarks

In this chapter, the predictive validity of the FP assessment was explored by correlating students' scores in the FP assessment with their scores in the FY academic courses. The findings revealed that proficiency in English is a moderate predictor of academic achievement in general. However, the strength of the predictive validity was found to vary according to students' self-evaluations and specialisations, but not their gender or college. The higher the students evaluated their language proficiency, the higher the FP assessment predictive validity became. The FP predictive validity was strong for the CS and English language groups,

moderate for the IT group and non-significant for the IBA group. Also, the predictive validity of high school assessment was investigated with regard to FY academic achievement: though scores in high school were strong predictors of the students' scores in FY assessment for the IT and CS groups, they were not good predictors for the IBA and English language (education) groups. The overall predictive validity of scores on the high school assessment was similar to the predictive validity of scores on the FP assessment.

Chapter 11: General Discussion and Conclusions

11.1. Introduction:

This chapter aims to provide an overarching discussion of the findings reported in Chapters 5, 6, 7, 8, 9 and 10. It will revisit the study methodology and discuss the importance of using different data sources. Then it will situate the findings in the context of related and comparable literature. The implications and limitations of this study are identified in the last section of the chapter.

The discussion of the results will triangulate evidence generated by various methods (i.e., document analysis, questionnaires, focus groups, interviews and correlation studies) in order to build a comprehensive argument about the effectiveness of the Foundation Programme (FP) assessment and its predictive validity. In evaluating language assessment, such a comprehensive approach that considers not only the product but also subsequent uses and interpretations of assessment scores is needed.

> The structure and purposes of assessment programs vary and therefore the evidence required to support the claims being made will vary, but validation always involves the evaluation of the proposed interpretations and uses of the assessment scores (Kane, 2011, p.10).

Validation of assessment instruments should consider both evidence and consequences of scores' uses and interpretations. Messick emphasises that both a *test*[19] and *performance assessment* are governed by similar validation principles and their evaluation should both include evidence and consequences as part of its validity argument.

> Hence, performance assessments must be evaluated by the same validity criteria, both evidential and consequential, as other assessment …

[19]He explains the difference in using the terms 'tests' and 'performance assessment' in a later paper saying "the current Educational reform movement in the USA puts considerable stock in the notion that performance assessments, as opposed to multiple-choice tests, will facilitate improved teaching and learning" (1996, p. 241)

> because they are not measurement principles, they are social values that have meaning and force outside of measurement wherever evaluative judgements and decisions are made. (1994, p. 13)

This chapter is divided into five main sections. The first section (11.2) discusses the effectiveness of FP assessment and builds an argument around the two threats to validity: construct underrepresentation and construct irrelevance (Messick, 1989). The second section (11.3) considers the findings on the predictive validity of FP assessment, and presents evidence from various instruments to offer a comprehensive picture, compare the findings of this study to those of previous studies, and identify factors influencing the predictive validity. The third section (11.4) summarises the findings on three central issues: criterion/norm referenced assessment, the test/performance (continuous) assessment distinction and the impact of FP assessment.

The last three sections of this chapter discuss the implications, limitations and recommendations of this study. The implications are categorised into theory, practice and policy in section (11.5). Then, the limitations of the focus, methods or findings of this study are identified, and possible contributors to the limitations are suggested in section (11.6). The last section (11.7) discusses a number of recommendations drawn from the findings of this study for future research on this topic.

11.2. The Effectiveness of FP Assessment

4. How well did the process of assessing students' English language performance, through continuous assessment and tests, function in the Foundation Programme?[20]

The answers found to this question are given in Chapters 5 to 9 using evidence from document analysis, questionnaires, teacher interviews, and student focus groups in both phases of the study. A central topic to the discussion will be the two threats to validity identified by Messick (1989) namely construct underrepresentation and construct irrelevance (see Section 3.2).

[20] The order of these questions is different from that presented in Chapter 1 to improve the clarity of the discussion, but the original numbers are kept.

11.2.1. Evidence from Document Analysis

1.1. What were the processes and procedures that were followed in writing and using the assessment instruments as depicted by the official documents?

This question was dealt with in Chapter 5. Document analysis strongly suggests that assessment in the General English Skills course (GES) and in the Academic English Skills course (AES) were seriously affected by the two validity threats. Instances of construct underrepresentation occurred in AES performance assessment: an analysis of the marking scale used to evaluate students' reports seemed to indicate that the procedures and technicalities of writing were more focused upon than the linguistic features of a written piece (see Section 5.3.2.2). This problematic issue was exacerbated by the documented difficulties in implementing the marking scales consistently. In marking AES assessment (i.e., the report and presentation), within and across colleges, inconsistency in implementing the marking scales was indicated in the standardisation and moderation documents (see section 5.3.3). This finding accords with a number of studies that generally criticise performance assessment for its low reliability (e.g., Clapham, 2000) which is caused by raters' inconsistency (e.g., Banjeree and Wall, 2006; Eckes, 2008; Elder, Barkhuisen, Knoch, & Randow, 2007) or other factors (e.g., Shohamy, 1995).

Another source of construct underrepresentation is the absence of assessment tasks on the reading and listening skills from the AES assessment. The AES course specifications document lists a number of listening and reading learning outcomes that are excluded from AES assessment, which only evaluate students' skills in writing and presentation. The AES course specifications and textbook are misleading with regards to what skills students are expected to master by the end of the course. The claim that a certain score in AES assessment gives information about students' proficiency in four language skills (i.e., reading, writing, listening and speaking) is untrue; in fact, it only conveys information about writing and speaking skills.

Elements of construct irrelevance are evident in both GES assessment and AES assessment. In the GES tests, some of the test tasks were of higher levels or different genres than the ones used in the text books, though, the difficulty level of GES learning outcomes and test tasks conform to each other (see Section 5.3.2). In the AES assessment, construct irrelevance was clear in the length and complexity of the report writing when compared to the short simple paragraphs supplemented in the AES text book (see Section 5.3.2.2). These features of both AES and GES assessment have led to "excess reliable variance" that increased difficulty and consequently signified construct irrelevance.

2.1. What were the national and international policies on teaching and assessing language that influence assessment in Oman? And how does FP assessment correspond to these policies?

The findings suggest that one of the main causes of construct irrelevance in GES and AES assessment is the adoption of national standards for the FP as learning outcomes without modifying them to accommodate the needs of the students, or match the content and level of the textbooks (see Section 5.3.4). The standards set by the Oman Academic Accreditation Authority (OAAA) were devised to ensure the quality of the national foundation programmes. Internationally, outcomes-based assessment is used to improve the quality, accountability and transparency of assessment (e.g., Brindley 2001; Llosa, 2007). These standards are also a means in globalised English medium higher education to standardise and control the quality of language preparation programmes (see Section 1.3.2). However, in this case, the standards were used not only as FP learning outcomes but as FP assessment specifications, creating a gap between stated outcomes and assessment specifications on one hand and the materials used on the other hand; this consequently increased the difficulty level of FP assessment.

Similarly, the AES writing task was based on the national standards, but its marking scale was not. The descriptors of the marking scale were of a lower difficulty level and of a different focus than the stated learning outcomes or national standards. The

findings suggest that students could obtain a pass mark in the AES report writing if it showed evidence of incorporating teachers' comments, originality, submitting on time, and being within the word limit (see Section 5.3.2). This finding supports the argument that negative consequences can be generated by implementing outcome-based assessment when developed at a distance from teachers and teaching contexts (Arkoudis & O'Loughlin, 2004; Mckay, 2007), and leads us to caution against direct and uncritical implementation as seems to have been the case in the FP.

11.2.2. Evidence from Students and Teachers in Phase 1

1.2. How was the reliability and validity of FP assessment viewed by students and teachers?

This question was addressed in Chapters 6, 7, 8 and 9 which presented the students' and teachers' views on FP assessment reliability and validity in the first and second phases. The views of these stakeholders constitute one aspect of the consequential basis of assessment validation (Fulcher, 1996, 2010). Besides, some writers feel that the impact of assessment on stakeholders should be considered when building validation arguments (e.g., Hamp-Lyons, 2000; Norris, 2008).

Chapter 6 displayed the students' and teachers' views on the reliability, validity and impact of FP assessment as expressed through the questionnaires in Phase 1. Generally, both the students and their teachers tended to view FP assessment positively, but had different perceptions of its reliability and validity. The students tended to respond to the items on FP validity and reliability more positively than did their teachers, and also rated its reliability higher than its validity, unlike their teachers. These views were clarified by the views expressed in focus groups and interviews presented in Chapter 7. The students expressed more concerns about the content and construct of FP assessment (i.e., difficulty levels or types of tasks) than they did about the inconsistency in using marking scales. On the other hand, the inconsistent use of the marking scales in AES assessment dominated in most of the teachers' interviews. The teachers' concerns about the reliability of using marking scales to evaluate students' writing and presentations mirrored a continuing theoretical debate on the adequacy of using performance assessment in high stakes

contexts considering its low reliability (e.g., Bachman & Palmer, 1996; Teasdale, 2000; Linn, 1993). The proponents of such assessment argue that its reliability should be considered differently because they are based on a different view of what language is (Gipps, 1994; Fox, 2008); besides, some studies have found that performance assessment correlated highly with reliable standardised measures (Llosa, 2007). The findings of this study show that the issues of validity and reliability of tests and performance assessment are not only debated in the theoretical sphere but also in practice by teachers and students using their own language. They also show that the quality of FP assessment is problematic; this will be elaborated on in the following section

Like the discussion of the findings generated by document analysis, that of teachers' and students' views in the first phase raises issues of FP irrelevance and construct underrepresentation. Firstly, construct irrelevance can be inferred from students' and teachers' views on three problems: the high difficulty levels of AES and GES assessment tasks; plagiarism in report writing; and, variability in using marking scales. Elements of GES tests and AES report writing required language mastery levels that were perceived to be higher than that of the students. In GES tests, the listening task, which constituted 20% of the test scores, presented the students with an unfamiliar listening genre. The grammar tasks, which constituted 10% of the test marks, were described as challenging and the students said that they were not prepared for these tasks. The reading task in the midterm test was described as difficult because of the unfamiliar topic. Besides, the limited time available for responding to the test tasks generally intensified the difficulty of GES tests according to students (see Section 7.2.2). Similarly, most of the teachers highlighted and criticised the high difficulty levels of GES test tasks. They added that the unavailability of past test papers had increased the uncertainties about, and difficulty of, the tests, and expressed a need for more assessment instruments of a better quality (see Section 7.3.2). It was felt that more regular assessment instruments would provide students with much needed feedback. Using tasks irrelevant to the assessed construct and thus increasing the difficulty is an element of construct-irrelevance

(Messick, 1989), the students' and teachers' views suggest that GES tests had this weakness.

Secondly, elements of construct underrepresentation are clear in the teachers' and students' views about AES assessment. Both teachers and students pointed out the unsuitability of the marking scale used for the writing assessment in AES in terms of its focus and use. Some teachers felt that the criteria were lower than they should be and allowed students who had not mastered the required writing skills to pass; this finding has also been reached from document analysis. Besides, most of the students and teachers raised a concern that the marking scales were used differently by different teachers. This corresponds with the findings reported in numerous studies about inconsistency in implementing marking scales (e.g., Brindley, 1998; Gipps, 1994; Hay & Macdonald, 2008). This study provides a different type of evidence on rater inconsistencies; one that is based on students' and teachers' perceptions. Most studies on this topic have explored raters' variability through the actual process of marking (e.g. Banjeree & Wall, 2006; Clapham, 2000; Eckes, 2008; Lumley, 2002). Banjeree & Wall (2006) argue that raters' inconsistency in using marking scales inevitably results in invalid assessment. In line with this argument, the analysis of stakeholders' views suggests that the nature of this "invalidity" can be seen as construct underrepresentation.

11.2.3. Evidence from Students and Teachers in Phase 2

2.2. What were the teacher and student perceptions of the assessment tools' effectiveness and their roles in shaping language assessment in retrospect?

This question was addressed in Chapters 8 and 9 in which the findings from Phase 2 student questionnaire/focus groups and teacher questionnaire/interviews were analysed and discussed. In this phase, the students were undertaking academic courses along with an English language course. The questionnaires revealed that most of the teachers were moderately satisfied with FP and FY English assessment. However, the students' responses showed a low level of satisfaction and often indicated their belief that their language levels were inadequate for FY study. About

70% of the students seemed to believe that English language assessment should be changed in both AES and GES courses. This change would not entail lowering the criteria, in fact, most of the students (all of whom scored more than 75% of the total FP marks or less than 55% of FP marks) disagreed with the statement that FP assessment criteria should be lowered. The nature of the desired change was clarified in focus groups in which most of the students seemed to feel that the FP should have been more intensive and its assessment should have been stricter and more challenging. In a number of cases, they stated that FP study was not taken seriously, and its assessment encouraged a lazy attitude towards FP study. Though the teachers' responses in the questionnaire suggested an overall satisfaction with the FP assessment, they also suggested a general feeling of the inadequacy of the students' language proficiency for FY academic study. In the interviews, these views were expanded on and a consensus on the lack of readiness of most students for FY study emerged. Also, there was an agreement that the FP assessment was ineffective because of its lenient criteria and students' lack of motivation.

The views of the students and teachers in the second phase about the leniency of the FP seem, on the surface, to contradict their views in the first phase. However, a closer look reveals that the students were referring to the FP study when they maintained that it should be stricter and more intensive, not the assessment itself. Similarly, the teachers criticised the FP assessment criteria, not the instruments themselves; they linked the inappropriateness of the students' language levels to the lack of clarity and leniency of the criteria used in the FP.

In the focus groups, the majority reported experiencing challenges in reading assigned materials, understanding lectures and expressing their views in writing and speaking. The teachers reiterated some of these difficulties and added others such as lacking essential study skills. Previous studies have identified similar non-linguistic difficulties as influencing factors on academic achievement (Hill, Stortch & Lynch, 1999). Xu (1991) argues that such difficulties (and others') are better predicated by exploring students' self-evaluations rather than their scores in language assessment. In the current study, the linguistic difficulties that the students faced in the FY (i.e.,

difficulties coping with the reading, listening, speaking and writing requirements) may indicate that these skills were underrepresented in FP assessment. The interpretation that passing the FP means being prepared for academic study in the FY - as could be inferred from the purpose statement of the FP - is arguably imprecise considering the serious linguistic difficulties that students face in the first semester of academic study.

4.2. How did teachers and students think language accuracy should be considered in assessing academic assignments?

This question was addressed in both Chapters 8 and 9. The students and teachers expressed uncertainties about including language accuracy as a criterion in the scale to evaluate written assignments in academic courses. Most of the students (68%) maintained that they did not know whether language accuracy and content were both assessed in the academic courses. In the focus groups, they explained that the marking scales for both the academic and English language courses were not clear, deadlines were not identified early enough and feedback was scarce (see section 7.2.1). In the interviews, the academic courses teachers revealed their awareness of the lack of clear criteria for marking written assignments and sometimes lack of consensus on marking criteria among teachers of a single course (see Section 7.3.1). Though there are specified marking scales in the English language course, the English language teachers rarely referred to them in their discussions; when they did, the scales were often described as lenient or unsuitable.

In similar higher education contexts where English is used as the medium of instruction for non-native speakers, similar uncertainties about how English language should be considered in assessing written assignments of academic courses have been reported among students and teachers. Al-Badwawi (2011) reported that the teachers in her study did not make any explicit reference to reported criteria by which students were supposed to be assessed, despite the fact that English language teachers were officially obliged to use a centrally developed scale. The uncertainties shared by the students and teachers are very likely to result in lowering the reliability

of assessment instruments and arguably increase its difficulty level for the students (Norton & Starfield, 1997).

The students expressed mixed views on whether language accuracy should be/is assessed in academic writing. Most students agreed that language accuracy should be considered in assessing written assignments in English language courses, but less so in the academic courses (see Section 8.2.3). Similar, but more varied, attitudes were detected from the teachers' responses; most supported the view that language accuracy should be a criterion in evaluating written assignments in academic courses, though most of them also believed that language inaccuracies should be overlooked when the general meaning is understood (see Section 8.3.3). Similar attitudes were found in the teacher interviews, but it seems that the academic teachers and English language teachers had different approaches to language accuracy in marking written assignments. Most of the academic teachers favoured overlooking language mistakes in written assignments when the overall meaning was comprehensible, whereas; the reported views of the English language teachers split three ways, some considered both language and content; others focused on one, or the other. A similar finding was reported by Al-Badwawi (2011, p.179) who found that academic teachers evaluated the content of a written piece but tended to disregard the linguistic features. Such a variance in use of specific criteria to evaluate written assignments is widely reported in the literature on academic writing in higher education (e.g., Lea & Street, 1998).

I will now try to interpret or reformulate these findings in Messick's terms. Messick (1989) claims that validity (1) is not a quality of tests or test scores, (2) is a matter of degree, (3) is linked to uses and interpretations, and (4) is a unitary concept. From this perspective, the FP assessment cannot be described as entirely invalid, but invalidity may be ascribed to specific uses and interpretations that evidentially or consequentially showed features of construct irrelevance and construct underrepresentation. For example, FP assessment was intended to be used as achievement assessment (see Section 5.3.1), but the evidence from document analysis indicated that it was based on the national GFP standards, not the course specifications and materials. Another example is the claim that passing the FP

assessment entailed being able to handle FY study with minimal linguistic challenges (see Section 3.4) which was contradicted by the subsequent linguistic difficulties the students faced according to their own and their teachers' views (see Section 9.4.2.). Taking the unitary theory of validity that argues for including assessment consequences as part of evidence on assessment validity, it can be claimed that elements of construct irrelevance and underrepresentation occur in FP assessment, and thus the effectiveness of the FP assessment is questionable.

1.7. In all the above, were there any significant differences between the views of the students' grouping by college, gender, age, self-evaluation and teachers' grouping by gender, college, age, nationality, teaching and assessment experiences?

Within the student sample, the frequency of responses that implied dissatisfaction with FP assessment varied considerably among the groupings by college, specialisation and self-evaluation. Sur students were more dissatisfied with FP assessment than Rustaq students, and International Business Administration (IBA) students showed most dissatisfaction with FP assessment followed by the Information Technology (IT) and Communications Studies (CS) groups. Likewise, in the self-evaluation groups, the dissatisfaction level with FP assessment was higher in the *very good* group, followed by the *excellent*, *average*, *good* and *weak* groups. In investigating the effectiveness and predictive validity of FP assessment, the students' specialisations and self-evaluations appeared to be influencing factors not only on students' opinions but also on the strength of the predictive validity, as discussed in the following section.

11.3. Evidence on the Predictive Validity of the Foundation Programme

3. What was the predictive validity of the English language assessment for student performance on the academic courses?

The second main area that this study investigated is the predictive validity of FP assessment. Predictive validity means the extent to which performance on an assessment instrument correlates with performance on another when the instruments

are administered with a time difference (Weir, 2005). Previous studies have reported inconsistent results on the predictive power of English language tests in higher education contexts (e.g., Graham, 1987; Bayliss, 2006; Elder, 1993; Cotton & Conrow, 1998). According to some of these studies, the strength of the correlation is influenced by several variables, and it is widely agreed that even international, reputable and validated English language tests (i.e., IELTS or TOEFL) should not be used as the sole criterion for admission into higher education; these tests have shown low to moderate predictive validity for academic achievement. Graham (1987, p. 561) writes that:

> The Educational Testing Service (1985, p.16) itself urges institutions not to use TOEFL scores as the sole basis for admission decisions, not to use rigid cut-off scores, and not to use the scores for predicting academic performance.

This issue will be further discussed under the impact of FP assessment in section 11.3.3. In the following two sections, the findings of this study on FP predictive validity will be discussed using evidence from a correlation study, document analysis and teachers' and students' views.

11.3.1. Correlation Study and Document Analysis

3.1. Did student performance in English language assessment in the FP correlate positively with their performance in academic courses?

The findings on this question were presented in Chapter 10. The predictive validity of FP assessment was moderate at 0.31, $p<0.01$. This means that students' proficiency in English does not highly correlate with their academic achievement. Actually, only 9% of students' academic achievement is explained by their English language proficiency. As has been argued throughout the previous chapters, proficiency in English and academic achievement are not expected to be highly correlated as each focuses on different skills. However, studying predictive validity is important for informing policies on using the English language as a gatekeeper and assisting in understanding the factors that affect academic achievement.

A number of studies on the predictive validity of internationally recognised tests (e.g., IELTS and TOEFL) or in-house tests have reached inconsistent conclusions (see Section 3.4). The strength of the correlation between language assessment and academic achievement reported in these studies ranged from non-significant (e.g., Kerstjen & Nery, 2000) to strong (e.g., Al-Musawi & Al-Ansari, 1999). Strong correlation coefficients have been usually reported in studies where English is not only the medium of instruction but also the subject of study.

3.2. Did the strength of correlation between the language proficiency and academic achievement differ significantly when students' scores in English language tests only or continuous assessment only were used, instead of the overall scores in both?

This question was addressed in Chapter 10 by correlating the students' average scores in AES and their average scores in GES with their average scores in the academic courses. It was found that the strength of the predictive validity was influenced by the type of assessment: the predictive validity was found to be higher for tests ($r = 0.367$, $p < 0.01$) than it was for the continuous assessment ($r = 0.272$, $p < 0.01$). Two possible reasons for the difference in the predictive validity of the two types of assessment are tentatively identified as inconsistent use of AES marking scales, and their lower in difficulty level compared to tests. As reported above, the students and teachers raised concerns about raters' variability and the teachers claimed that the writing marking scales were allowing students to pass too easily.

3.3. Did the groupings by college, gender, self-evaluation and specialisation show significant differences in correlations between language proficiency and academic achievement?

Though gender did not significantly affect the predictive validity, other factors such as specialisation and self-evaluation did. For instance, the predictive validity of FP assessment was stronger for the CS group ($r= 0.64$, $p < 0.01$) and English Language (Education) group ($r= 0.57$, $p < 0.01$) than it was for the IT ($r = 0.41$, $p < 0.05$) and IBA ($r = 0.18$, $p = 0.12$) groups. It seemed that assessment in some disciplines required more command of the English language; thus, students' proficiency in English was more likely to affect their academic achievement.

The findings of the current study confirms some previous findings on the influence that such variables (e.g., academic disciplines and self-evaluations) had on the power of predictive validity. A number of studies have identified academic disciplines as an influencing factor; students' English language proficiency levels correlated more highly with their academic achievement in specific disciplines than in others (Davies, 1990; Al-Musawi & Al-Ansari, 1999; Elder, 1993; Cotton & Conrow, 1998; Huong, 2000; Lynch, 2000). Huong notes that some disciplines are linguistically more demanding, but the nature of this linguistic demand has not been explained or explored by previous studies (see Section 3.4). Similarly, students' self-evaluation was another variable that affected the strength of the predictive validity; the predictive validity of FP assessment was higher for the groups with higher self-evaluations. Very few studies have highlighted the evident role that students' self-evaluation can have on the strength of the predictive validity of English language assessment and in anticipating difficulties of academic study (Xu, 2000). It should be noted here that it should not be expected in this type of study that the predictive validity will be high because the assessment instruments correlated (i.e., English language assessment and academic assessment) actually evaluate distinct constructs (i.e., English language proficiency and academic achievement). Therefore, as some writers (e.g., Lynch, 2000) have argued, the English language assessment used as qualifying or placement instruments to higher education are actually fulfilling their purposes regardless of their low to moderate predictive validity. Nevertheless, the importance of understanding the factors that influence academic achievement in the first year of higher education along with proficiency in English language is vital to the improvement of higher education.

In chapter 5, the role played by language proficiency in academic achievement was investigated further by analysing the linguistic requirements of course specifications and test materials of three main academic disciplines: IT, IBA and CS. As has been mentioned earlier, several studies have found a marked difference in the strength of the predictive validity of English language assessment between the academic disciplines, but hardly any study has investigated this difference further or looked

into the nature of the linguistic demands of these disciplines. In the current study, the linguistic demands of the IT, IBA and CS courses were investigated by analysing the course specifications, assessment tasks and test papers. The analysis of these documents helped to explain the findings on the predictive validity of FP assessment. It was found that IBA and IT coursework and test tasks required limited command of the English language and utilised questions that demanded defining concepts or reciting learned material, whereas CS coursework and test tasks made greater demands on students' proficiency in English language as they involved writing long essays, presenting topics or arguing for and against propositions. Understanding the nature of the linguistic requirements of the academic courses assessment could assist in identifying the skills that should be taught in the FP and understanding language assessment predictive validity further. Such research has been called for as a way forward in predictive validity studies (Davies, 1990; Fox 2004), and this thesis can claim some originality in this area.

11.3.2. Predictive Validity in Students' and Teachers' Perceptions

3. How did stakeholders understand the relationship between the student performances in the English language assessment and their performances in the academic courses' assessment?

This question was partly addressed through questionnaires in Chapters 6 and Chapter 8. In the first phase of this study, both students and teachers said that most of the students would pass FP assessment but were very likely to struggle in academic study due to their inadequate language skills. In the second phase, most of the students and their teachers felt that proficiency in English language had an influence on academic achievement. This view was not surprising; several previous studies have emphasised that academic performance is enhanced by higher language proficiency levels especially in linguistically demanding courses (e.g., Woodrow, 2006; Powers, Kim, & Wang, 2008) and this association is more evident below a certain threshold (Phillips, 1987) .

What were students' and teachers' perceptions of the importance of the predictive validity?

This question was responded to in Chapters 7 and 9. In focus groups, most of the students from Sur College maintained that proficiency in English language played a major role in academic achievement, whereas in Rustaq College most of the students seemed to believe that it had a lesser role along with other factors, and some seemed to feel that it had none. Rustaq students' general view seemed to be that the importance of language proficiency in academic achievement was dependent on the extent to which test tasks in their academic subjects required mastery of the language, reproducing memorised information and application of practical skills.

Like students, most teachers tended to consider the association between language proficiency and academic achievement as a strong one and regarded English language as an important criterion in assessing written assignments. In Chapter 9, the findings from the teacher interviews confirmed their view, as expressed in the questionnaire, that there is a relationship between English language proficiency and academic achievement. Teachers' emphasis on the strength of this relationship differed between the colleges and among specialisations. Sur teachers emphasised the importance of the role of English language in academic achievement more than Rustaq teachers. Actually, two of the teachers from Rustaq (from IT and IBA departments) felt that proficiency in English language was irrelevant to academic success. Given that most of the participants in the interviews from Sur College were English language teachers and CS teachers, it is understandable that proficiency in English language was considered to be of high importance in academic achievement, particularly when considering the findings on predictive validity (see Section 10.3).

11.4. Related Topics

Three topics central to the questions of this study that were briefly touched upon in the previous discussion of the results, will now be discussed in this section. These are

criterion/norm-referenced assessment; preference of Test/CA; and, FP assessment impact.

11.4.1. Criterion/Norm-Referenced Assessment

1.6. What types of assessment (criterion/norm-referencing) were used? And how?

This section addresses the assessment design aspect of the question, as the marking aspect was discussed above. It has been pointed out previously that, though CAS academic regulations and OAAA standards required using criterion-referenced assessment, document analysis revealed that FP assessment followed norm-referenced procedures in analysing the GES test scores (see Section 5.3.1). This section provides a detailed discussion of the implications of using norm-referenced assessment. Policy makers' preference for criterion-referenced assessment has been explained by the need to ensure that a certain standard is met by all concerned institutions (Brindley, 2001; Llosa, 2007; Sizmur and Sainsbury, 1997). The distinction between criterion-referenced and norm-referenced assessment is usually made at the administrative levels (Martuza, 1977), but the borderline between them tends to be obscured in classroom practices and sometimes teachers use one instead of the other. The documents analysed in this study reflected this uncertainty and confusion between stated policies that explicitly mandated using criterion-referenced tests and actual or described practices that used norm-referenced analytical procedures. This finding may be explained by the fact that norm-referencing procedures have been used in educational systems for much longer than the criterion-referenced ones.

> Over the years, standard procedures for testing and measurement within a norm-referenced framework have become well known to educators (Hambleton, Swaminathan, Algina, & Douglas, 1978).

Hughes believes that textbooks in language assessment also have a role in the prevalence of norm-referencing procedures in testing.

> Books on language testing have tended to give advice that is more appropriate to norm referenced testing. One reason for this may be that the procedures for use with norm-referenced tests … are well established, while those for criterion-referenced tests are not (2003, pp. 21-22).

All this means that teachers were, and perhaps are still, more accustomed to implementing norm-referenced analytical procedures than the less common criterion-referenced ones, and this may help explain why GES tests writing and scores analysis mainly, followed norm-referencing procedures, instead of criterion-referencing ones.

Given that norm-referenced assessment aims at evaluating test takers' performances against each other rather than a certain set of attainments, the difficulty level of tasks included is bound to the students' performances, not to predefined outcomes such as those mandated by the OAAA. Therefore, passing GES tests depends on how well the students perform against each other. Any interpretations based on using criterion-referenced assessment assume mastering a set of outcomes, and the same applies to an outcome based assessment, but these interpretations are invalid for the actual FP assessment, as the GES test follows a norm-referenced model, not a criterion-referenced one.

Though the GES tests were basically norm-referenced, AES assessment closely mirrored the learning outcomes stated by the OAAA and followed a criterion-referenced model. The AES course learning outcomes and assessment specifications correspond in level and focus to those stated by the GFP standards (2009), but the learning materials targeted a lower level of English language proficiency. The discrepancy in levels between taught materials and assessment tasks increased the difficulty of AES assessment and arguably has led to plagiarism which threatens assessment validity (Fulcher, 2010). The findings indicate that norm-referenced assessment is not suitable for use as a gatekeeper for higher education in contexts where specific outcomes should be mastered. Also, implementing criterion-referenced assessment should take care not to be so occupied by the outcomes that

the content of the course is forgotten. The course outcomes and test specifications should not only reflect national standards but also textbooks and materials used.

11.4.2. Tests /CA

> 1.4. What were the differences between the 'continuous assessment' model used in the Academic English Skills course and the ‘test’ model used in the General English Skills course in terms of effectiveness, accuracy, and preferences of teachers and students?

In discussing the effectiveness of FP assessment, both the teachers and the students distinguished in their views between GES tests and AES assessment (i.e., writing a report and conducting a presentation). Most teachers did not seem to believe that standardised tests were more reliable or valid, and preferred using performance assessment - the students had similar views. Also, the teachers generally believed that learning happens in performance assessment, a view which has been advocated in the assessment literature (e.g., Broadfoot, 2003).

11.4.3. The Impact of FP Assessment

> 1.3. How was the impact of FP assessment perceived by students and teachers?

Assessment consequences, including the interpretations made based upon assessment scores, are another aspect of validity that should be considered as evidence in validation studies (Fulcher & Davidson, 2007). This has been touched upon in discussing the findings on the effectiveness and predictive validity of FP assessment, but this section focuses on the value implications attached to assessment scores and their use.

> The unified concept of validity integrates considerations of content, criteria, and consequences into a construct framework for testing rational hypotheses about theoretically relevant relationships. These hypotheses relate to data patterns expected not only on the basis of provisional score meaning but on the basis of value implications of the score interpretations, and on the basis of presumed relative import of intended and unintended outcomes of score use. (Messick, 1989, p. 8)

The assessment impact involves all consequences of assessment in the classroom context and beyond, including any value implications attached to the interpretation of the test scores. This study investigated the social and political aspects. Given the high stakes of FP assessment, the status of English language in Oman higher education, the purpose of FP assessment (Shohamy, 2001; Shohamy, 1996; Ross, 2008), it was expected that social impact (e.g., perception of self, effects on social life, sense of fairness) and political impact (e.g., FP assessments as gatekeeper to higher education, scores in English language assessment affecting students' employability) would be prominent in both the students' and teachers' perceptions. In the student interviews, very few students associated failing the FP with negative social connotations. The majority of students and teachers only seemed concerned about the political impact of English language assessment, that is, students' access to higher education and graduates' access to the labour market.

One of the significant differences between genders was found in the students' responses on the impact of proficiency in language on their future careers and national and international policies of the country. Female students expressed more agreement with the items that implied a political impact of English language assessment than did the male students. A similar attitude towards the impact of English language assessment was reflected in the female teachers' responses. Though identifying gender differences in FP assessment was not the main focus of the study, the revealed differences can perhaps be understood in the context of previous research on gender differences in Omani society. The female participants view of the importance of English language skills for future careers could be tentatively explained by the low employability rate of women compared to men in Oman as well as in other Gulf countries (Klasen & Lamanna, 2009), and by the extra challenges women face when attempting to attain middle to upper management positions (Al-Lamky, 2007). Gender differences were not a core element of this study; however, this appeared to be an influencing factor in participants' perceptions of the impact of FP assessment.

11.5. Implications

Whilst acknowledging the limitations of this study, I believe that its findings can contribute to the wider literature on English language assessment in higher education contexts. I hope the findings will be useful to language teachers, programme designers and policy makers involved in the FP in Oman as well as to the general population of language testers. It is also hoped the they will not only contribute to practical aspects of language assessment but also to theoretical ones, as the two following sections discuss.

11.5.1. Theoretical Issues

This section discusses two main theoretical issues relating to validity theory and more generally to the predictive validity of language assessment. In 1955, Cronbach and Meehl introduced the concept of construct validity in psychological and educational tests. Fulcher and Davidson (2007, p. 181) argue that Cronbach and Meehl's paper planted the seeds for the later view of the unified concept of validity as suggested by Messick (1998) and the subsequent frameworks for validation arguments. In Messick's view (1989, p.5), validity is "an inductive summary of both existing evidence for and the actual as well as the potential consquences of score intrepretations and use". This unified view of validity is pragmatic. Fulcher and Davison (2007), in arguing for this pragmatic view, say:

> What we learn from different approaches and definitions of validity is that validity theory is changing and evolving … our understanding of the validity of test use for a particular purpose is dependent upon evidence that supports that use (p.18).

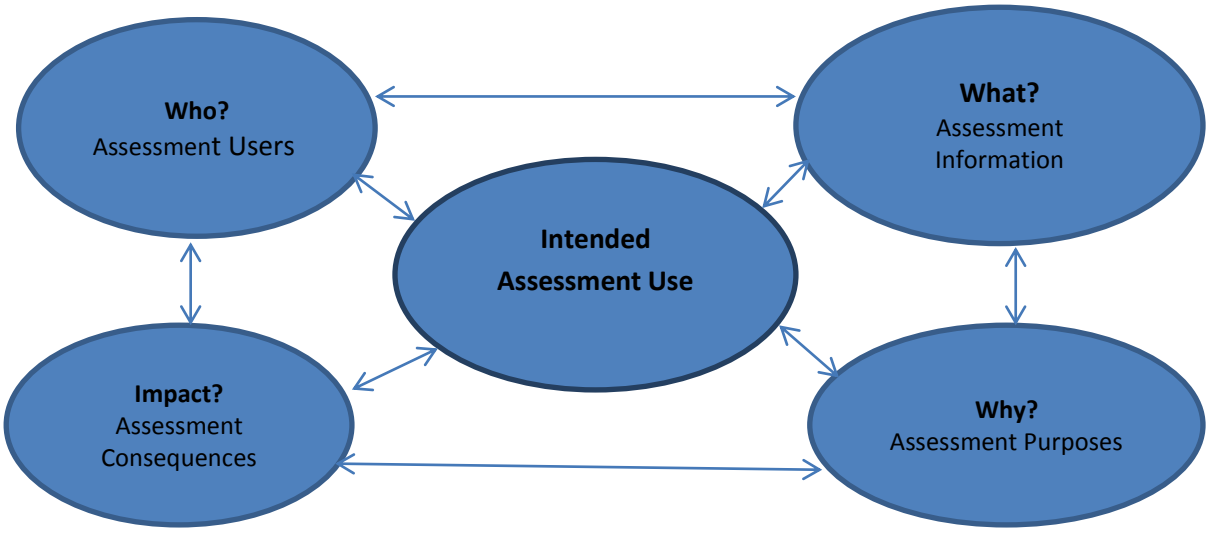
The range of evidence Fulcher and Davidson talk about is very wide and variant depending on uses, intrepertations, or purposes of a validator. Norris (2008) raises concerns about the impreciseness of educational assessment validation, he argues that Messick's unified approach to test validation is very general and imprecise about what to include in a validation process and how to conduct it. He says that it lacks structure and focus, and instead he proposes adopting programme evaluation approaches to organise and lead the process of assessment validation. Four steps are required to engage in validity evaluation as Norris suggests: (1) treatment of educational assessment as a programme not an instrument, (2) identifying reasons

and purposes for conducting validity evaluation, (3) prioritising purposes, (4) identifying suitable methods. Norris summarises the rationale for shifting from traditional validation to validity evaluation saying:

> the shift to validity evaluation seeks to transform validation into a worthwhile and relevant endeavour by making its purposes explicit and by conceptualising its use within a specific community with clearly defined interests in a particular assessment programme (2008, p.76).

During the course of this study, it became evident that a study of the effectiveness of the FP assessment programme necessitated investigating assessment as a comprehensive programme that included multiple variables as constructs such as: curriculum, stakeholders, uses and consquences. I found that studying the validity of an assessment inspired by the unified concept of validity lacked focus and that Norris's model is more procedural and thus gives more concrete guidance to validators. This model also assists in focusing the purposes of an assessment programme validation, identifying varient contributing elements to the programme and providing informative information for future improvements. Recasting validation as validity evaluation following this model entails reconceiveing validity as an "educationally relevant concept rather than a preoccupation of psychometricians" and requires considering the purposes, models and methods of programme evaluation. Therefore, I believe that using a model such as the one suggested by Norris (2008) shown in figure 11.1 would assist not only in structuring the validation process but also in identifying and prioritising the purposes of the evaluation and focal areas of interest.

Figure. 11.1. Specification of Intended Assessment Use (Norris, 2008, p.102)

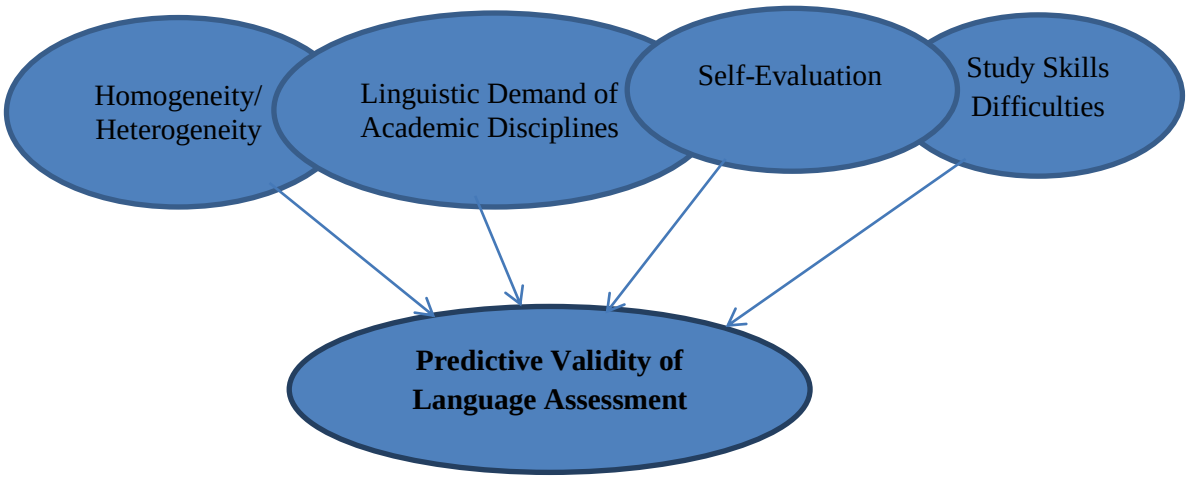


The above model identifies four areas that should be studied, and provides evidence on validation evaluation: assessment purposes, assessment information, assessment users and assessment consquennces. Collecting evidence on these areas focuses the process of validation and accounts not only for validity aspects as suggested by Messick but other elements of an assessment programme as well, such as: purposes and users. Considering these areas presents assessment programmes as comprehensive programmes and indicates that their implementation entails ramifications.

The second theoritical issue of this study is how to study the predictive validity of language assessment. Traditionally, the predictive validity of an assessment instrument has been mainly studied through correlating scores in two instruments without much attention being paid to the contextual factors. In this study, some factors that affect the strength of the predicitve validity were investigated. Some of these factors seem to have a clear role in this study as in previous studies. These include: the sample's homogeneity/heterogeneity, academic dicipline, and self-evaluation. The predictive validity of language assesmsent seems to be stronger in homogenous samples, and linguistically demanding diciplines (e.g., CS). The role of other variables, such as study skills difficulties, remains unclear.

The following diagram, shows suggested factors to be investigated when conducting predictive validity studies for their possible influence on the strength of the predictive validity of language assessment.

Figure 11.2. Factors to be Explored in Studying the Predictive Validity of Language Assessment



The purpose of proposing this model is to emphasise that these factors should be considered in future research. Previous studies, separately, indicated the role of one or more of these factors in influencing the strength of the predictive validity of language assessment (see Section 3.4). Understanding these factors will provide more insight into the nature of the predictive validity of language assessment in academic achievement.

Both qualitative and quantitative methods could be used in such an investigation of the role of contextual factors in the predictive validity of English language assessment with regards to academic achievement. Qualitative exploratory research can be used to determine the variables included in each factor, after that, factor analysis can be used to reduce these variables and build up scales for each of the factors. The findings on these scales can be then analysed using multiple regression to find out the predictive power of each of these factors.

11.5.2. Practical Implications

The findings of this study have a number of practical implications for English language teaching and assessment in CAS. Evidence from various parts of the study supports the implications listed below:

- It seems that there is a mismatch between the levels of teaching materials used and test tasks. This mismatch seems to be a result of adopting the national GFP standards irrespective of the students' levels and materials used. This issue should be dealt with by revising the FP curriculum (i.e., objectives, teaching materials and assessment) to correspond more closely with both the students' levels and the standards. One way this could be achieved is by replacing currently used textbooks for ones of more advanced levels that conform with the national standards and provide students with the needed challenge that FP materials currently lack [students frequently mentioned the FP study was not challenging enough for them while the assessment instruments were].

- The GES tests clearly follow norm-referencing procedures in the way test item analysis is used. This should be changed to be criterion-referenced, first to conform to stated policies in this regard and second to enhance the validity of FP assessment by bringing it closer to course specifications and actual teaching.

- A number of concerns were raised about the AES writing marking scale such as its leniency and inconsistent use by teachers. The descriptors of the marking scales should be changed to better reflect the outcomes in the AES course specifications. Though standardisation and moderation procedures of marking are documented in the assessment policies of the English language department at CAS, these procedures seem not to be fully implemented. More effort should be made to implement these procedures.

- A related finding to the previous one is that the AES assessment showed a lower value of its predictive validity than did the GES tests. This finding should be considered in borderline cases where students' scores are very close to the cut-off point (50 out of 100). The present practice is that if a students' score is 48 or 49, it is rounded up to 50 (i.e., the passing score). I recommend that in such cases, students' scores in the GES assessment should be given more weight. This recommendation also supports the current policy followed in the FP of allowing admitted students to take a challenge exam (i.e., an English test offered to those who do well in the placement test which if they pass will permit them to undertake FY courses without undergoing performance assessment tasks or taking FP English language courses).

- There are uncertainties about the assessment content and structure. Both the students and teachers said that they were uncertain about specific aspects of FP assessment such as: marks distribution, specifications of instruments and even descriptors of marking scales. Assessment details should be shared with both the students and teachers at the beginning of the academic semester to eliminate any underperformance due to uncertainties and increase the validity of FP assessment.

- The students shared that they received a lack of feedback that could be attributed to the summative nature of FP assessment. Instead, formative assessment instruments that provide enough feedback to students and that show a high degree of validity should be considered. [Both students and teachers were asking for more assessment instruments as a means for extra feedback]. This could be achieved by incorporating smaller units of classroom assessment early in the semester to allow enough time for feedback. These units should be validated prior to use and teachers should be trained to mark them as consistently as possible, preferably using a similar marking scale to that used for other performance assessment tasks.

- One issue that arose in the second phase of this study is the lack of a common scale for marking written assignments in the academic courses. The role of English language as a criterion in marking written assignments seemed to be unclear and left to the judgement of teachers, most of whom tended to ignore language inaccuracies when meaning was clear. Students also seemed unclear about the role of English language accuracy in marking academic written pieces. This possibly generates invalid assessment, and therefore it is suggested that the role of language accuracy in assessing written assignments should be made clear in College policies for both teachers and students.

- Plagiarism is an issue of concern to students and teachers. More support and training on what plagiarism is and how to avoid it should be given to the students, and clearer guidelines about how to deal with it should be given to teachers.

11.5.3. Policy Implications

The findings of this study can feed into national educational policies in three ways. First, the results show that students recognised that their performance in English language assessment had a major impact in terms of access to the labour market and higher education. With such a high-stakes assessment, its validity should be taken seriously to ensure that assessment uses and interpretations are supported by theoretical rationales and empirical evidence. Decisions linked with youth higher education opportunities or job opportunities are very critical and should be based on valid information. The findings of this study reveal moderate to low predictive validity of English language assessment with regards to academic achievement, but students' proficiency in English language plays a major role in accessing Omani higher education. Considering the findings of this study and other comparable ones, it is recommended that in admission to higher education, proficiency in English language should be considered as a criterion along with students' academic achievement, but used differently. Currently, higher education programmes that use English as a medium of instruction require a certain level of achievement in high school English language courses equal to that required in academic courses (Higher

Education Admission Centre, 2012, p.83). Instead, if a high school graduate obtained the academic achievement level required but not the English language level, he still should be considered for higher education admission but not if he meets the language requirement but not the academic one.

The second policy implication relates to the finding that some academic disciplines are more dependent on linguistic proficiency than others. Academic achievement in disciplines such as Communication Studies is strongly correlated with English language proficiency. Therefore, it is recommended, subject to supporting evidence from studies with larger samples across a wider spectrum of higher education institutions, that the cut-off point used in the FP should become higher for the CS discipline and others similarly dependent on English language proficiency.

The third policy implication is related to allocating more time to equipping students with appropriate study skills given the reported difficulties in this area which students face in the First Year of academic study. Both teachers and students identified a number of non-linguistic challenges as barriers to academic achievement. It is vital to ensure students' mastery of these skills before or during academic study, for example, by introducing a specific course that deals with them in the Foundation Programme. In designing such a course a needs analysis should be conducted before planning the curriculum.

11.6. Limitations

Though this study, it is hoped, has contributed substantial findings on English language assessment in an EFL context, it is limited in several ways. Firstly, because the topic was about assessment, data collection was conducted at the end of two academic semesters; this was intended to give the students and teachers ample opportunity to become familiar with the assessment system and instruments specifications. The choice of this period of time, however, made many teachers and students reluctant to participate in focus groups and interviews due to being engaged in assessment related tasks; the students had to submit reports and prepare for speaking tests while the teachers had a large amount of marking and preparation to

do for the final exam. This was unfortunate, but with the type of summative assessment used in CAS, collecting data earlier was not possible.

Secondly, the study used a convenient sample that depended on students' and teachers' willingness to participate. Though the sample came from two of the six Colleges of Applied Sciences, it did not represent the entire population of CAS students. Future research with a larger sample is needed to validate the findings of this study.

The third limitation of this study comes from the four questionnaires used in both phases. Though the questionnaires were piloted and Cronbach Alpha and Inter-item correlation were tested, the questionnaires still showed some flaws (e.g., the wording of some items and low Cronbach Alpha for certain topics). The items included under each topic were not always re-statements of the same construct; sometimes they were addressing different aspects of one construct. Also, I noticed that using a Likert scale that included five categories, one of them denoted "no opinion", resulted in a large number of the participants selecting this option. In certain items, the number of participants who selected "no opinion" was 50% or more. One solution is to use a graded Likert scale that does not have a middle (no opinion) point.

Furthermore, using focus groups in this study as one of the methods posed some challenges. These challenges included training participants in this method of sharing views, re-coding all opinions, transcribing and analysing generated data. However, the two main challenges were the "no shows" which is a common and documented disadvantage of this method (Bryman, 2004), and ensuring free and equal expression of views. My impression was that, though all participants were encouraged to be involved in the discussion, some of them preferred to listen. Also, sometimes one opinion dominated the discussion and overrode others. Any decision to use focus groups should be an informed understanding of all of their advantages and disadvantages.

11.7. Recommendations for Future Research

There are a number of recommendations that can be made based on the findings of this research. In studying the predictive validity of assessment instruments, further investigation of the factors influencing the strength of predictive validity is needed. The findings of this study, similar to previous comparable ones, revealed that the homogeneity of the sample, students' self-evaluations, academic disciplines and types of assessment instruments all influenced the predictive validity of the FP assessment. For future research on language assessment predictive validity with regards to academic achievement, considering these factors might provide a clearer picture of the role of language in academic achievement and explain some of the variances in findings reported by previous studies (see section 11.4.1).

In the Omani context, further research is needed on the role of English language in determining access to both higher education and the labour market. Omani youths' proficiency in English language seems to have a strong impact on their future, but very few studies have investigated the nature this impact. It could be argued that this is also adversely affecting the human resources of the country in which only those who can use English language with a certain level of proficiency are allowed access to higher education and high-ranking jobs, and the majority of high school graduates, who cannot, are thus excluded. More studies should be conducted on the impact of the present policies on this group.

11.8. Concluding Remarks

English language assessment plays a critical role in Omani higher education and its impact is evident in recent student protests. The present study found problems with the validity of the Foundation Programme assessment including construct-irrelevance and construct-underrepresentation, which may help explain students' frustrations. The high stakes of FP assessment in CAS necessitate urgent action on the areas of doubtful validity.

The study also investigated the correlation between proficiency in language and academic achievement (i.e., predictive validity). It was found that the predictive validity of FP assessment is only moderate and varied depending on students' specialisations, types of assessment instrument, and self-evaluations. Analysing samples of tests from three different specialisations suggested that some specialisations required more command of English language than others. This may explain the different levels of FP predictive validity for students from different specialisations and assist in future decisions for amending the cut-off point required for academic study in higher education institutions. It also confirms the findings of previous studies suggesting that proficiency in English language had a rather limited role in academic achievement and shows that other factors need to be explored.

In general the findings of the study indicated that the students seemed to be to some extent successful in the FY academic studies despite the language difficulties they faced. Also, FY assessment received overall satisfaction from the participants in terms of content and scales specifically in terms of the relatedness between what is taught and what is tested as well as introducing academic language skills (i.e., specific academic vocabulary and academic writing). There are good lessons that should be drawn from FY assessment to be implemented in FP assessment.

Reference List

Al Badwawi, H. (2011). *The Perceptions and practices of first year students' academic writing at the Colleges of Applied Sciences (*Unpublished PhD thesis). The University of Leeds, Leeds.

Al Bandary, M. S. (2005). Meeting the challenges: The development of quality assurance in Oman's Colleges of Education. *Higher Education, 50*, 181-195.

Alderson, C. J. (2009). *The Politics of Language Education.* Bristol: Multiligual Matters.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation* . Cambridge: Cambridge University Press.

Al-Issa, A. (2005). The implications of the teacher educator's ideological role for the English languge teaching system in Oman. *Teaching Education, 16*(4), 337-348.

Al-Issa, A. (2006). The cultural and econimic politicas of English langauge teaching in the Sultanate of Oman. *Asian EFL Journal Quarterly*. *8* (1) 194-218.

Al Kharusi, H. (2008). Effects of classroom assessment practices on students' achievment goals. *Educational Assessment, 13*, 243-266.

Al-Lamki, S. M. (1998). Barriers to Omanization in the private sector: The perceptions of Omani graduates. *The International Journal of Human Resources Mangement, 9*(2), 378-400.

Al-Lamki, S. M. (2002). Higher education in the Sultante of Oman: The challenge of access, equity and privatization. *Journal of Higher Education Policy and Mangement, 24*(1), 76-86.

Al-Lamki, S. M. (2006). The development of private higher education in Oman. *International Journal of Private Education, 1*, 54-77.

AL-Lamky, A. (2007). Feminizing leadership in Arab societies:The perspectives of Omani female leaders. *Women In Management Review, 22*(1), 49 - 67.

Allwright, J and Banerjee, J, 1997, *Investigating the accuracy of admissions criteria: A case study ina British university*, Institute for English Language Education, Lancaster University, Lancaster.

Al-Mahrooqi, R. (2012). English communication skills: How are they tought at schools and universities in Oman? *English Langugae Teaching, 5*(4), 124-130.

Al-Musawi, N. M., & Al-Ansari, S. H. (1999). Test of English as a Foreign Language and First Certificate of English Tests as Predictors of academic sucess for undergraduate students at the University of Bahrain. *System, 27*, 389-399.

Al-Rahbi, I. A. (2008). *An Empirical study of the key knowledge economy factors for sustainable economic development in Oman*(unpuplished PhD thesis). Victoria University, Melbourne.

Al Shemli, S. H. (2009, September). Higher education in the Sultanate of Oman: Planing in the context of globalization. Paper presented at *IIEP Policy Forum*. UNESCO: Paris.

Alsarimi, A. M. (2001). New trends in assessment in the Sultanate of Oman: Goals and characteristics. *Educational Measrment: Issues and Practice*, 27-29.

Altbach, P. G., & Knight, J. (2007). The internationalization of higher education: Motivaion and realities. *Journal of Studies in International Education, 11*(3), 290-305.

Altheide, D. L. (1996). *Qualitative Media Analysis*. Thousand Oaks, Calif: Sage.

Arkoudis, S., & O'Loughlin, K. (2004). Tensions between validity and outcome: Teacher assessment of written work of recently arrived immigrant ESL students. *Language Testing, 21*(3), 284-304.

Ashworth, P., Bannister, P., & Thorne, P. (1997). Guilty in whose eyes? University students' perceptions of cheating and plagiarism in academic work and assessment. *Studies in Higher Education, 22*(2), 187-203.

Atkinson, P. & Coffey, A. (2004). Analyzing documentary realities. In D. Silverman (Ed.), *Qualitative research*. (pp. 77-90). London: Sage Publications.

Bachman, L. F. (1990). *Fundemental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F. (2002). Some reflections on task-based langauge performnace assessment. *Langauge Testing, 19*(4), 453-476.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Bachman, L., & Plamer, A. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Langauge assessment in practice.* Oxford: Oxford University Press.

Baker, D. (1989). *Language testing: A critical survey and practical guide.* London: Edward Arnold.

Ball, S. J. (1998). Big policies/small world: An introduction to international perspectives in education policy. *Comparative Education, 34*(2), 119-130.

Banjeree, J., & Wall, D. (2006). Assessing and resporting perfromances on pre-sessional EAP courses: Developing a final assessment checklist and investigating its validity. *Journal of English for Academic Purposes, 5*, 50-69.

Bangert-Drowns, R. L., Kulik, C.-L. C., & Kulik, J. A. (1991). The Instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213-238.

Bayliss, A. (2006). IELTS as a predictor of academic langauge performnace. Paper presented at the Australian International Education Coference *(AIEC)*, Perth. Abstract retrieved from http://www.aiec.idp.com/pdf/BaylissIngram%20(Paper)%20Wed%201630%20MR5.pdf

Beretta, A. (1992). Evaluation of language education: An overview. In J. C. Alderson, & A. Beretta (Eds.), *Evaluating Second Language Education* (pp. 5-24). Cambridge: Cambridge University Press.

Bhola, H. S. (2003). Social and Cultural Contexts of Educational Evaluation: A global perspective. In T. Kellaghan, D. L. Stufflebeam, & L. A. Wingate, *International Handbook of Educational Evaluation.* Dordrecht: Kluwer Academic Publishers.

Black, P. (2003). Assessment for Learning- Lessons from Research and from Practice. Paper presented in *Byond Exams conference,* Bristol.

Blanche, P. (1989). Self-Assessment of Foreign-Learning Skills: Implications for teachers and researchers. *language learning, 39*(3), 313-339.

Bowen, G. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*. *9* (2), 27 – 40.

Breeze , R., & Miller, P. (2008). Predictive validity of the IELTS listening test as an indicator of student coping ability in Spain. *IELTS Research Reports. 12*, 1-34.

Briggs, R. S., & Cheek, M. J. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality, 54*, 475-89.

Brindley, G. (1998). Outcomes-based assessmentand reporting in language learning programmes: A review of the issues. *Langauge Testing, 15*(1), 45-85.

Brindley, G. (2001). Outcomes-based assessment in practice: some examples and emerging insights. *Language Testing, 18,* 393-407.
Brindley, G. (2003). Classroom-based assessment. In D. Nunan (ed.). *Practical English Language Teaching.* New York: McGraw Hill.
Broadfoot, P. (2007). *An Introduction to Assessment* . New York: Continuum.
Brown, A. (1995). The Effect of rater variables in the development of an occupational specific language performance test. *Language Testing, 12*(1), 1-15.
Brown, F. G. (1976). *Principles of Educational and Psychological Testing.* New York: Holt, Rinehart and Winston.
Brown, G. T., & Hirschfeld, G. H. (2008). Students' conceptions of assessment: Links to outcomes. *Assessment in Education: Policy and practice, 15*(1), 3-17.
Brown, J. D. (1990). *Testing in Language Programs.* New Jersey: Prentice Hall Regent.
Brown, J. D. (1996). *Testing in Language Programs(2nd ed.).* New Jersey: Prentice Ahll Regents.
Brown, J.D. and Hudson, T. (1998). The alternatives in language assessment. TESOL Quarterly 32, 653–75.
Bryman, A. (1984). The debate about quantitative and qualitative research: Aquestion of method or espistimology? *The British Journal of Sociology, 35*(1), 75-92.
Bryman, A. (2004). *Social Research Methods.* USA: Oxford University Press.
Bryman, A. (2006a). Integrating quantitative and qualitative research: How it is done? *Qualitative Research, 6*(1), 97-113.
Bryman, A. (2006b). Paradigm peace and the implications for quality. *International Journal of Social Research Methodology, 9*(2), 111-126.
Bryman, A. (2008). Advances in mixed-methods research: Theories and applications. London: Sage Publications.
Cammara, W. J., & Kimmel, E. W. (2004). *Choosing Students: Higher education admission tools for the 21st century.* NewJersay: Lawrence Erlabum Associates.
Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. Richards, & R. Schmidt (Eds.), *Language and Communication* (pp. 2-27). London: Longman.
Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second langauge teaching and testing. *Applied Linguistics, 1*, 1-47.
CAS. (2009). *English Department Assessment Handbook.* Muscat: Ministry of Higher Education.
CAS. (2010a). *Colleges of Applied Siences Regulations.* Muscat: Ministry of Higher Education.
CAS. (2010b). *Course Specifications for Foundation English* . Muscat: Ministry of Higher Education.
CAS. (2010c). *English Department Handbook.* Muscat: Ministry of Higher Education.
CAS. (2010d). *Foundation Program 2010-2011.* Muscat: Ministry of Higher Education.
CAS. (2011). *Assessmnet Policies.* Muscat: Ministry of Higher Education.
Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254-272.
Cheng, L. (1999). Changing assessment: WashBack on teacher perceptions and actions. *Teaching and Teacher Education, 15*, 253-271.
Cheng, L., Andrews, S., & Yu, Y. (2011). Impact and consequences of school-based assessment (SBA): Students' and parents' views of SBA in Hong Kong. *Language Testing, 28*(2), 221–249.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performnace: Some evidence from American Universities. *Langauge Testing, 29*(3), 421-442.

Clapham, C. (2000). Assessment and testing. *Annual Review of Applied Linguistics, 20*, 147-161.

Cope, N. (2011). Evaluating locally-developed language testing. *Australian Review of Apllied Linguistics, 34*(1), 40-58.

Cotton, F., & Conrow, F. (1998). An investigation of the predictive validity of IELTS amongest a group of international students studying at the University of Tasmania. *IELTS Research Reports, 1*, 72-115.

Creswell, J. (2003). *Research design: Qualitative, quantitative and mixed methods approaches (2nd ed.).* Thousand Oaks, CA: SAGE Publications.

Creswell, J. W. (2011). Controversies in mixed method research. In N. Denzin, & Y. Lincoln (Eds.), *The SAGE Handbook of Qualitative Research* (pp. 269-283). London: SAGE.

Creswell, J. W., & Miller , G. A. (1997). Research methodologies and the doctoral process. *New directions for Higher Education, 99*, 33-46.

Creswell, J. W., & Miller, L. D. (2010). Determining validity in qualitative enquiry. *Theory Into Practice, 39*(3), 124-130.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Crotty, M. (1998). *The Foundations of Social Research: Meaning and perspective in the research process.* Los Angeles: Sage.

Currie, P. (1998). Staying out of trouble:Apparent plagirism and acaemic survival. *Journal of Second Language Writing, 7*(1), 1-18.

Dale, R. (1999). Specifying globalization effects on national policy: A focus on the Mechanisim. *Journal of Education Policy*, 1-17.

Dancey, C. P., & Reidy, J. (2004). *Statistics Without Maths for Psychology.* Harlow: Peason .

Darlaston-Jones, D. (2007). Making connections: The relationship between epstimology and research methods. *The Australian Community Psychologist, 19*(1), 1927.

Davidson, F., Turner, C. E., & Huhta, A. (1997). Language testing standards. In C. Clapham, & D. Carson, *Language Testing and Assessmsent* (Vol. 7, pp. 303-311). Holland: Kluwer: Dordrecht.

Davies, A. (1990). *Principles of Language Testing.* Oxford: Blackwell.

Davies, A. (2008a). *Assessing Academic English : Testing English proficiency, 1950-1989(the IELTS solution).* Cambridge: Cambridge University Press.

Davies, A. (2008b). Ethics, professtioalism, rights and codes. In D. Alan, *Encyclopedia of Language and Education*. pp. 424-444.

Davies, A. (2012). Kane validity and soundness. *Langauge Testing, 29*(1), 37-42.

Davies, A., & Elder, C. (2005). Validity and validation in langauge testing. In E. Hinkle, *Handbook of Research in Second Language Teaching and Learning* (pp. 795-813). Mahwan, NJ: Lawrence Erlbaum.

Donn, G., & Al Manthri, Y. (2010). *Globalization and Higher Education in the Arab Gulf States.* Oxford: Symposium Books.

Donn, G., & Issan, S. (2007). Higher education in transition: Gender and change in the Sultanate of Oman. *Scottish Educational Review, 39*(2), 173-185.

Dornyei, Z. (2007). *Reesrach Methods in Applied Linguistics and Mixed Methodologies.* Oxford: Oxford University Press .

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.
Elder, C. (1993). Language proficiency as a predictor of performance in teacher education. *Melbourne Papers in Langauge Testing, 2*(1), 68-89.
Elder, C., Barkhuizen, G., Knoch, U., & Randow, J. V. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, 37-64.
English Department. (2011). Instructions for Report Writing and Presenting in AES Course. CAS.
Erling, E. J., & Hilgendorf, S. K. (2006). Language policies in the context of German higher education. *Language Policy, 5*, 267-292.
Fairclough, N. (2003). *Analyzing Discourse: Textual analysis for social research.* London &New York: Rutledge.
Feast, V. (2002). The impact of IELTS scores on performance at University. *International Education Journal, 3*(4), 70-85.
Fielding, J., & Gilbert, N. (2006). *Understanding Social Satistics.* London: Sage Publications.
Forster, N. (1994). The Analysis of Company Documentation. In C. Cassell and G. Symon (eds). Qualitative Methods in Organizational Research. London: Sage.
Fox, J. (2004). Test decisions over time: Tracking validity. *Language Testing, 21*(4), 437-465.
Fox, J. (2008). Alternative assessment. In E. Shohamy, & N. H. Hornberger, *Ecyclopedia of Language and Education: Language Testing and Assessment* (pp. 97-109). New York: Springer.
Fulcher , G. (2010). *Practical Language Testing.* London: Hodder Education.
Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing*, 23-51.
Fulcher, G., & Davidson, F. (2007). *Language Testing And Assessment: An advanced resource book.* London: Routledge.
Gipps, C. (1994). *Beyond Testing.* London: The Falmer Press.
Gipps, C. (1994). Developments in educational assessment: What makes a good test? *Assessment in Education, 1*(3), 283291.
Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24*, 355-392.
Glen A. Bowen. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27-40.
Graham, J. J. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly, 21*(3), 505-521.
Green, A. (2003). *Test impact and English for academic purposes: Acomparative study in backwash between IELTS preparation and university presessional courses* (Unpuplished PhD thesis). University of Surrey, Roehampton.
Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university presessional langauge courses. *Assessment in Education: Principles, Policy and Practice, 14*(1), 75-97.
Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy, 24*(1), 23-37.
Guba, E. G., & Lincoln, Y. S. (1989). *Fourth Evaluation Generation.* California: Sage Publications.

Guilherme, M. (2007). English as a global language and education for cosopolitian citizenShip. *Language and Intercultural Communication, 7*(1), 72-90.
Hambelton, R. K., Swaminathan, H., Algina, J., & Douglas, B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48*(1), 1-47.
Hamilton, L. (2003). Assessment as a policy tool. *Review of Research in Education, 27*, 25-68.
Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In Hamp-Lyons, L., editor, *Assessing second language writing in academic contexts*. Norwood: Ablex, 241–76.
Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing*, 14, 295-302.
Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System, 28*, 579-591.
Harrison, Andrew. (1983). A Language Testing Handbook. In Roger H. Flavell *Essential Language Teaching Series*. London: The Macmillan Press
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.
Hay, P., & Macdonald, D. (2008). (Mis)appropriations of criteria and standards-referenced assessment in a performnace based subject. *Assessment in Education, 15*(2), 153-168.
Higher Education Admission Centre. (2010). *Student Guide to Admission in Higher Education Institutions.* Muscat.
Hill, K., Storch, N., & Brian, L. (1999). A comparison between IELTS and TOEFL as predictors of academic success. *IELTS Research Reports.2* (3). 62-73. Retrieved from http://www.ielts.org/PDF/Vol2_Report3.pdf
Hammersley, M. (2002). The relationsip between qualitative and quantitative research: Paradigm loyalty versus methodological eclecticism. In J. T. Richardson, *Handbook of qualitative research methods for pyschology and the social sciences* (pp. 159-174). Leicester: The British Psychology Society.
Huong, t. (2001). The predictive validity of the international English language test system (IELTS). *The Post Graduate Journal of Education Research, 2*(1), 66-96.
Hughes, A. (1986). A pragmatic approach to criterion-referenced foreing language testing. In M. Portal, *Innovations in language testing* (pp. 31-40). Windsor: NFER-NELSON.
Hughes, A. (2003).*Testing English for Language Teachers*. Cambridge. UK: Cambridge University Press.
Hymes, D. (1972). On communicative competence. In J. B. Pride, & J. Holmes, *Sociolinguistics.* England: Penguin Books.
Iowa State University. (2010). Can You Call IT a Focus Group?. Retrieved from http://www.extension.iastate.edu/publications/pm1969a.pdf
Jacob, B., & Levitt, S. (2002, August). Rotten apples: An investigation of the prevelance and predictors of teacher cheating. *The Quarterly Journal of Economics*. 118(3), pp. 843-77
Jochems , W. (1991). Effects of learning and teaching in a foreign language. *European Journal of Engineering Education, 16*(4), 309-316.
Jochems, W., Sinppe, J., Jan Smid, H., & Verweij, A. (1996). The academic progress of foreign students: Study achievement and study behaviour. *Higher Education, 31*(3), 325-340.

Johnson, J., & Lim, G. S. (2009). The influence of rater language background on writing perfromance assessment. *Langugae Testing, 26*(4), 485-505.
Kane, M. (2011, April). Validating score inerpretations and uses: Messick lecture. A paper presented in Langauge Testing Research Colloquium, Cambridg. *Language Testing, 29*(1), 3-17.
Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.
Kerstjen, M., & Nery, C. (2000). Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance. *IELTS Research Reports*, 86-108.
Kiely, R. (2001). Classroom evaluation- values Intrests and teacher developmet. *language Teaching Research, 5*(3), 241-256.
Kiely, R. (2009). Small answers to big question: Learning from language program evaluation. *Language Teaching Research, 13*(1), 99-116.
Klasen, S., & Lamanna, F. (2009). The Impact of gender inequality in education and employment on economic growth: New evidence for a panel of countries. *Femenist Economics, 15*(3), 91-132.
Knoch , U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304.
Langridge, R., J. Christian-Smith, and K. A. Lohse. 2006. Access and resilience: analyzing the construction of social resilience to the threat of water scarcity. Ecology and Society 11(2): 18. www.ecologyandsociety.org/vol11/iss2/art18/.
Lea , M., & Street, B. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education, 23*(2), 157-172.
Linn, R. L. (1993). Educational assessment: Extended expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(1), 1-16.
Linn, R., Miller, M. (2005) Measurement and Assessment in Teaching. New Jersey: Pearson.
Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing,* 24(4), 489-515.
Llosa, L. (2011). Standards-based classroom assessments of English proficiency: A review of issues, current developments, and future directions for research. *Laguage Testing,* 28 (3), 367-382.

Lorena, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing,* 489-515.
Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276.
Lynch, B.K. (2001). Rethinking assessment from a critical prespective. *Language Testing, 18*(4), 351-372.
Lynch, B. K. (1996). *Langauge Program Evaluation: Theory and Practice.* Cambridge: Cambridge University Press.
Lynch, T. (2000). An evaluation of the revised test of English at matriculation at the University of Edinburgh. *Edinburgh Working Papers in Applied Linguistics, 10,* 61-71.
Maleki, A., & Zangani, E. (2007). A survey on the relationship between English Llngauge proficiency and the academic achievement of Iranian EFL students. *Asian EFL Journal, 9*(1), 86-96.

Markus, K. (1998). Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible? *Social Indicators Research, 45*(1), 7-34.
Martuza, V. R. (1977). *Applying Norm-Referenced and Criterion-Referenced Measurement in Education.* Boston: Allyan and Bacon.
Maxwell, J. A. (1992). Understanding and Validity in Qualitative Research. *Harvard Educational Review*. 62 (3). pp. 279-301.
McKay, P., & Brindley, G. (2007). Educational reform and ESL assesmsent in Australia: New roles and new tensions. *Language Assessment Quarterly, 4*(1), 69-84.
McNamara, T. (2008). Language testing. In A. Davies, & C. Elder (Eds.), *The Handbook of Linguistics* (pp. 763-782). Malden: Blackwell Publishing.
McNamara, T. F. (1996). *Measuring second language performance.* London and New York: Addison Wesley Longman.
Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(5), 5-11.
Messick, S. (1994a). The interplay and consquences in the validation of performance assessments. *Educational Researcher, 23*(13), 13-25.
Messick, S. (1994b). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.
Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement*. *14* (4). 5-8
Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 24-1256.
Miles, M. B., & Huberman, A. M. (1994). Making good sense. In M. B. Matthew (Ed.), *Qualitative data analysis: An expanded sourcebook* (pp. 245-287). london: Sage.
Ministry of Higher Education. (2010a). *Colleges of Applied Sciences: Prospectus 2010-2011.* Muscat: Ministry of Higher Education.
Ministry of Higher Education. (2010b). *Executive by Law of Royal Decree 62/2007 Regulating the Colleges of Applied Sciences.* Muscat: Internatio0nal Printing Press.
Ministry of Higher Education. (2010). *Higher Education Admission Statistics for the Academic Year 2009/2010.* Muscat: Al-Karamel International for Media Services.
Mishler, E.G. (1990). Validation in inquiry-guided research: The role of exemplars in narrative studies. *Harvard Educational Review*. 60 (4). 415-443.
Nitko, A. J. (1995). Curriculum-based continuous assessment: A framework for concepts,. *Assessment in Education, 2*, 321-337.
Norris, J. M. (2006). The why (and how) of assessing student learning outcomes in college foreign language programs. *The Modern Language Journal*, 576-583.
Norris, J. M. (2008). *Validity evaluation in language assessment.* Frankfurt: Peter Lang.
Norris, J. M. (2009). Understanding and imrpoving languge education through program evaluation: Introduction to the special issue. *Langauge Teaching Research, 13*(1), 7-13.
Norton, B., & Starfield, S. (1997). Covert language assessment in academic writing. *Langauge Testing*, 278-294.
Nunan, D. (2003). The impact of english as a global language on educational policies and practices in the Asia-Pacific region. *TESOL Quarterly, 37*(4), 589-613.
O'loughlin, K. (2011). The interpretation and use of Proficiency test scores in university selection: How valid and ethical are they? *Language Assessment Qurterly, 8*, 146-160.

OAAA. (2009). *Oman academic standards for general foundation programs.* Muscat: OAAA.

Onwuegbuzie, A. J., & Leach, N. L. (2005). On becoming a pragmatic researcher:The importance of combining quantitative and qualitative research methodologies. *International Journal of Research Methodology, 8*(5), 375-387.

Owen, J. M. (2007). *Program evaluation: Forms and appraoches.* New York: The Guilford Press.

Pallant, J. (2007). *SPSS survival manual.* New York: Open University Press.

Pecorari, D. (2003). Good and original: Plagirism and patchwriting in academic second language writing. *Journal of Second Language Writing, 12*, 317-345.

Pennycook, A. (1994). *The cultural politics of English as an international langauge*. New York: Longman.

Pennycook, A. (1999). Development, culture and language: Ethical concerns in a postcolonial world(paper presented at the Fourth International Conference of Langauge and Development). Vietnam.

Philip, E. M. (2011). The effects of language anxiety on students' oral test performance and attitudes. *The Modern Language Journal, 76*(1), 14-26.

Phillips, D. (1987). Langauge proficiency assesmsnet and tertiary entry for non-English speaking students. *Journal of Tertiary Educational Adminstration, 9*(1), 77-88.

Phillips, N. & Brown, J.L. (1993). Analyzing communication in and around organizations: A critical hermeneutic Approach. *Academy of Management Journal*. 36 (6). 1547-1576

Phillipson , R. (1992). *Linguistic Imperialisim.* Oxford: Oxford Univeristy Press.

Philipson, R. (2010). The politics and the personal in language education: the state of which art. *Language and Education, 24*(2), 151-166.

Philpot, S. (2006). *New Headway Academic Skills:Level 2 student's book.* Oxford: Oxford University Press.

Powers , D. E., Kim, H. J., & Weng, V. Z. (2008). The redesigned TOEIC (lisening and reading) test: Relations to test-taker perceptions of proficiency in english. *ETS Research Report n.RR-08-56*.

Powers, D. E., Schedl, M. A., Leung, S. W., & Butler, F. A. (1999). Validating the revised Test of Spoken English against a ariterion of communicative sucess. *Language Testing, 16*(4), 399-425.

Rea-Dickins, P. (1994). Evaluation and English langauge teaching. *Langauge Teaching*, 71-91.

Rea-Dickins, P. (1997). So, why do we need relationships with stakeholders in language testing? A view from the UK. *Language Testing, 14*(3), 304-314.

Rea-Dickins, P. (2007). Classroom-based assessment: Possibilities and pitfalls. In J. Cummins, & C. Davidson, *International Handbook of English Language Teaching* (pp. 505-520). Springer.

Richards, L., & Morse, J. M. (2007). Coding. In L. Richards (Ed.), *Read me first for a user's guide to qualitative methods* (pp. 133-151). London: Sage.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*(1), 1-20.

Ross, S. J. (2008). Language testing in Asia: Evolution, innovation, and policy challenges. *Language Testing, 25*(1), 5-13.

Runte, R. (1998). Impact of centeralized examinations on teacher professionalism. *The Canadian Journal of Education, 23*(2), 166-181.

Scriven, M. (2003). Evaluation theory and methodology. In T. Kellaghan, D. L. Stufflebeam, & L. A. Wingate (Eds.), *International Handbook of Educational Evaluation* (pp. 15-30). Dordrecht: Kluwer Academic Publishers.

Seelen, L. P. (2002). Is performance in English as a second language a releveant criterion for admission to an English meduim university. *Higher Education, 44*(2), 213-232.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguisitcs, 15*, 188-211.

Shohamy, E. (2001a). Democratic assessment as an alternative. *Langauge Testing, 18*(4), 373-391.

Shohamy, E. (2001b). *The power of tests.* Harlow: Longman.

Shohamy, E. (2006). *Language policy: Hidden agendas and appraches.* New York: Routledge.

Shohamy, E. (2007). Language tests as language policy tools. *Assessment in Education, 14*(1), 117-130.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*, 298-317.

Sizmur, S., & Sainsbury, M. (1997). Criterion referencing and the meaning of national curriculum assessment. *British Journal of Educational Studies, 45*(2), 123-140.

Soars, L., & Soars, J. (2006). *New Headway Plus: Intermediate student's book.* Oxford: Oxford University Press.

Spolsky, B. (1995). *Measured words.* Oxford: Oxford University Press.

Spolsky, B. (2004). *Language policy.* Cambridge: Cambridge University Press.

Spolsky, Bernard. (2008). Language assessment in historical and future perspective. In Elana Shohamy & Nancy Hornberger (Eds.), Encyclopedia of language and education (Second ed., Vol. 7: *Language testing and assessment*, pp. 445-454). New York: Springer Science.

Stansfield, C. W., & Hewitt, W. E. (2005). Examining the predictive validity of a screening test for court interpreters. *Language Testing, 22*(4), 438-462.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models and applications.* San Francisco: Jossy-Bass.

Tashakkori, A. (2003). *Handbook of mixed methods in social and behavioural Research.* California: Sage Publications.

Teasdale A. and Leung, C. (2000). Teacher assessment and psychometric theory: a case of paradigm crossing? *Language Testing, 17*, 2, 163–184.

Torrance, H., & Proyr, J. (1998). *Investigating formative assessment: Teaching, learning and assessment in the classroom.* Buckingham: Open University Press.

Toulmin, E. S. (2003). *The uses of argument.* Cambridge: Cambridge University Press.

Vaus, D. (2002). Surveys in social research. St. Leonards, NSW : Allen and Unwin.

Vinke, A., & Jochems, w. (1993). English proficiency and academic sucess in international postgraduate eductaion. *Higher Education, 26*, 275-285.

Wall, D., & Alderson, C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*(1), 41-69.

Walliman, N. (2005). *Your research project.* London: Sage Publications.

Weir, C., (1993). *Understanding and developing language tests*. Hemel Hempstead: Prentice Hall.

Weir, C. (2005). *Language testing and validation. An evidence-based approach.*UK: Palgrave MacMillan.

Weir, C. J. & Green, A. (2002) *The impact of IELTS on the preparation classroom: stakeholder attitudes and practices as a response to test task demands.* Unpublished IELTS Research Report.(The British Council).

Wiles, R. , Heath, G., Crow, S. & Charles, V. (2008). The management of confidentiality and anonymity in social research. *International Journal of Social Research Methodology*. 11(5). 417-428.

Woodrow, L. (2006). Academic success of international postgraduate education students and the role of English proficiency. *University of Sydney Papers in TESOL, 1*, 51-70.

Xu, M. (1991). The impact of English langauge proficiency on international graduate students' percieved academic difficulty. *Research in Higher Education, 32*(5), 557-569.

Yen, D. A., & Kuzma, J. (2009). Higher IELTS score, higher academic performance? The validity of IELTS in predicting the academic performance of Chinese students. *Worcester Journal of Learning and Teaching*, 1-7.

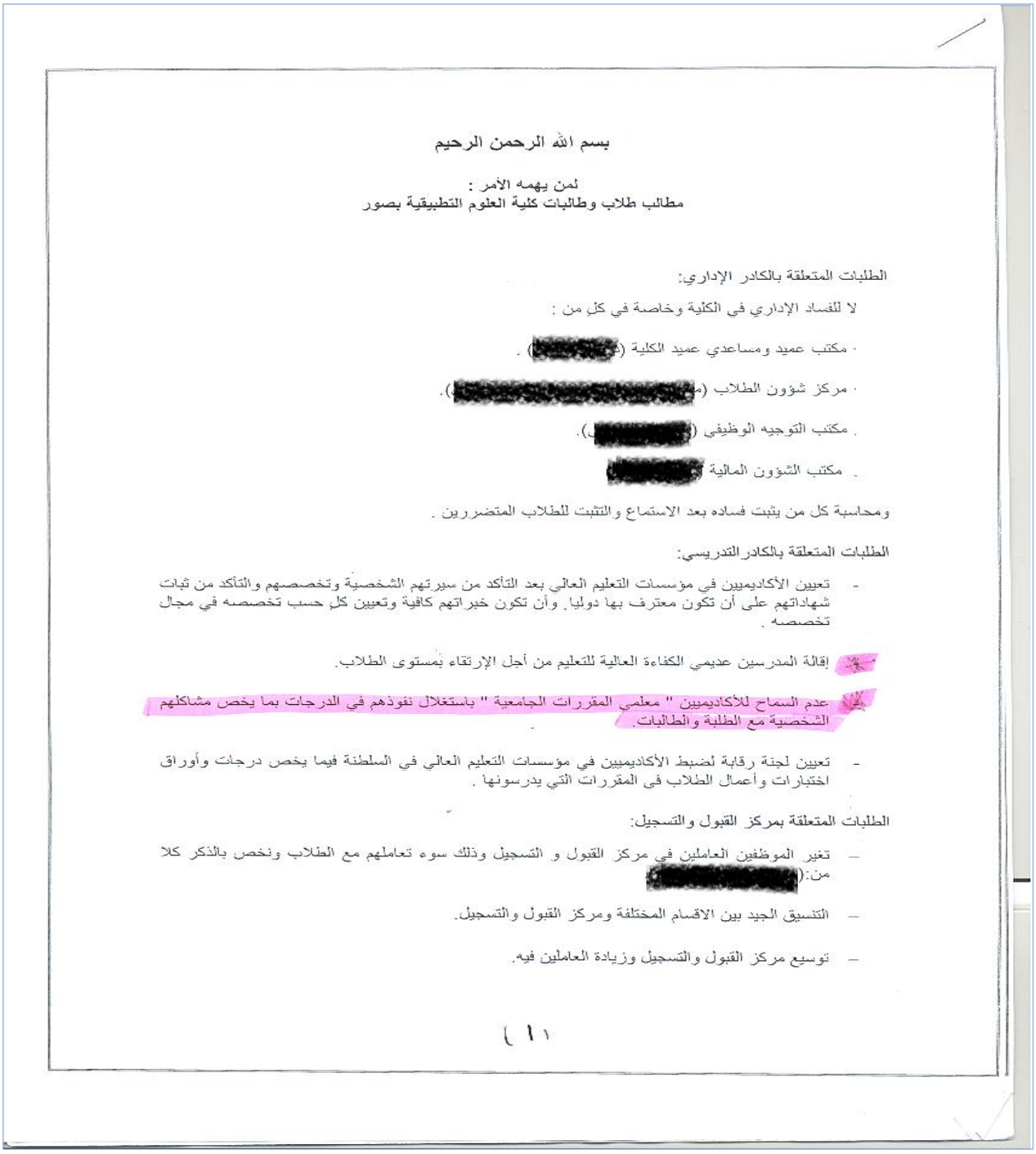
Yin, R. K. (2003). *Case study reseacrh: Design and method.* London: SAGE.

Yorke, M. (2003). Formative assessmant in higher education: Moves towards theory and enhancement of pedagogic practice. *Higher Education, 45*(4), 477-501.

Zabihi, R. (2011). Personality in English proficeincy and achievement. *Continental Journal of Education Research, 4*(1), 1-6.

Appendices

Appendix1.1. A flyer distributed in a student demonstration at Sur College in March 2011. The highlighted points concerning CAS assessment are translated in a following box.



بسم الله الرحمن الرحيم

لمن يهمه الأمر :
مطالب طلاب وطالبات كلية العلوم التطبيقية بصور

الطلبات المتعلقة بالكادر الإداري:

لا للفساد الإداري في الكلية وخاصة في كل من :

- مكتب عميد ومساعدي عميد الكلية ([redacted]) .
- مركز شؤون الطلاب ([redacted]).
- مكتب التوجيه الوظيفي ([redacted]).
- مكتب الشؤون المالية ([redacted])

ومحاسبة كل من يثبت فساده بعد الاستماع والتثبت للطلاب المتضررين .

الطلبات المتعلقة بالكادر التدريسي:

- تعيين الأكاديميين في مؤسسات التعليم العالي بعد التأكد من سيرتهم الشخصية وتخصصهم والتأكد من ثبات شهاداتهم على أن تكون معترف بها دوليا. وأن تكون خبراتهم كافية وتعيين كل حسب تخصصه في مجال تخصصه .

- إقالة المدرسين عديمي الكفاءة العالية للتعليم من أجل الإرتقاء بمستوى الطلاب.

- عدم السماح للأكاديميين " معلمي المقررات الجامعية " باستغلال نفوذهم في الدرجات بما يخص مشاكلهم الشخصية مع الطلبة والطالبات.

- تعيين لجنة رقابة لضبط الأكاديميين في مؤسسات التعليم العالي في السلطنة فيما يخص درجات وأوراق اختبارات وأعمال الطلاب في المقررات التي يدرسونها .

الطلبات المتعلقة بمركز القبول والتسجيل:

- تغير الموظفين العاملين في مركز القبول و التسجيل وذلك سوء تعاملهم مع الطلاب ونخص بالذكر كلا من:([redacted])
- التنسيق الجيد بين الاقسام المختلفة ومركز القبول والتسجيل.
- توسيع مركز القبول والتسجيل وزيادة العاملين فيه.

(١)

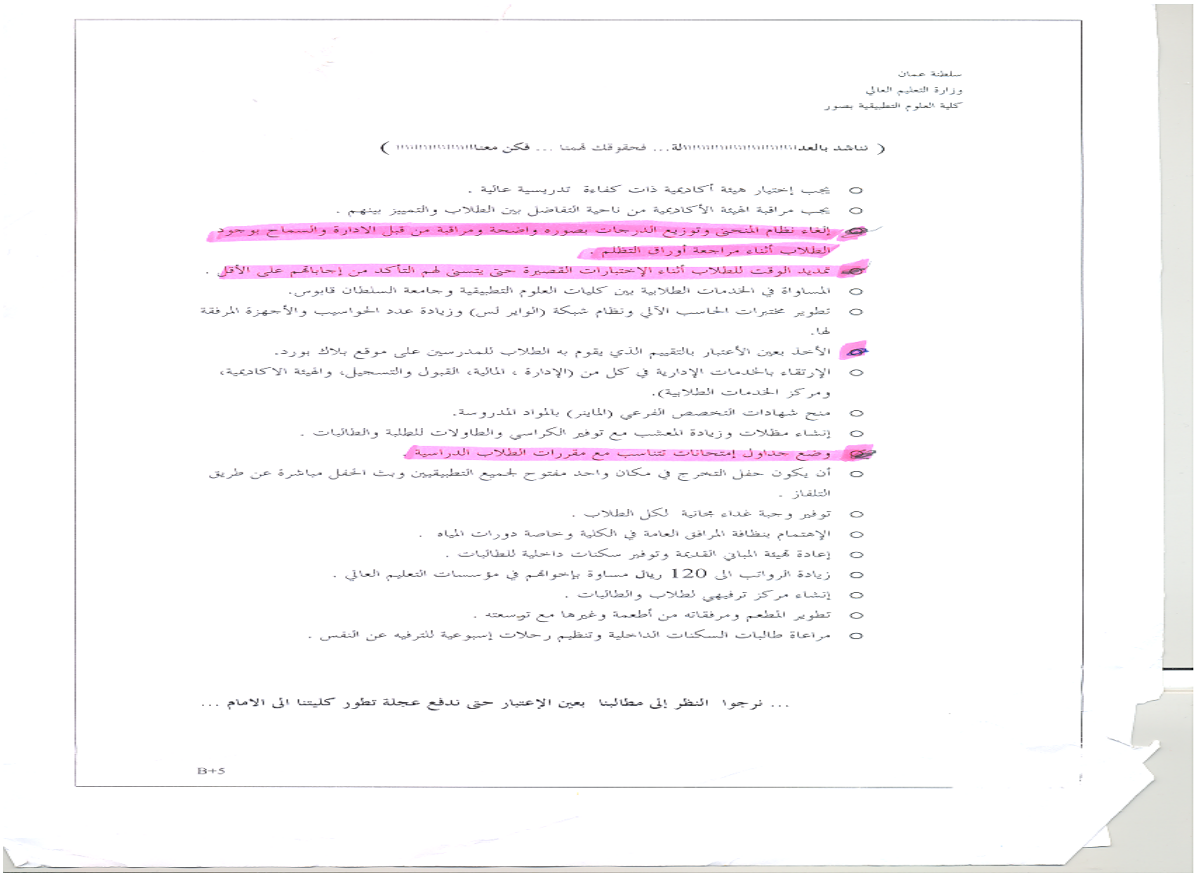
الطلبات المتعلقة بالحياة الأكاديمية للطلبة والطالبات:

- تشكيل اتحاد طلابي يتم انتخابة من قبل طلاب الكلية لمناقشة مشاكل الطلبة مع ادارة الكلية.
- تغير مسميات وطريقة احتساب المستويات بحيث يحصل الطالب على تقدير جيد من معدل 2.3 .
- في حالة رفع الطالب حالة تظلم يحق للطالب مراجعة ورقة اختباره ومناقشة الإجابات مع دكتور المادة " المقرر " بوجود لجنة تظلم خاصة بذلك.
- اعتماد مواد المايز ويعطى الطالب شهادة تثبت ذلك، وفي حالة تعذر ذلك يلغى المايز.
- تخفيض المقررات في جميع التخصصات وتحديد عدد معين من المحاضرات التي تدخل في الامتحانات النهائية.
- تعديل نظام الابتعاث في الكلية لتكملة الماجستير وذلك بتغير الشروط المطلوبة وجعل المعدل لا يقل عن 2.7 بدلا من 3.0 مثل قانون وزارة القوى العاملة.
- إعطاء الطلاب أكثر من ثلاث فرص في حالة تعرضهم إلى النزول تحت الملاحظة الاكاديمية و إرجاع جميع الطلبة الذين تعرضوا للطرد بسبب معدلاتهم و إعطاهم فصل واحد للطلبة الحاصلين على معدل اكثر من 1.5 ، وجعل معدل تحت الملاحظة يبدأ من ١.٨.
- ادخال نظام الدبلوم للراغبين في اخذ الشهادة و للراغبين في اكمال الدراسه يسمح لهم بذلك دون اي شروط او قيود تتعلق بالمعدل طالما الطالب فوق الملاحظة الاكادمية.
- السماح لأي طالب بالإلتحاق بالفصل الصيفي في حالة رغبتة في ذلك.
- السماح لطلبة الخريجين استكمال الدراسة الصيفية في حالة التخلف عن باقي الخريجين في مقرر او مقررين.
- الاعتراف بكل التخصصات في وزارة القوى العاملة.

الطلبات المتعلقة بالتوجيه الأكاديمي والوظيفي:

- تخصيص كادر اداري و توجيهي يهتم بإرشاد الطلاب في مرحلة تسجيل المقررات الدراسية.
- توفير فرص تدريب لجميع التخصصات بغض النظر عن السنة الدراسية وإعطاء الطالب حرية اختيار المكان المراد اخذ التدريب فيه.
- عمل تطبيق عملي وذلك بتخصيص ايام من الاسبوع للتدريب داخل احدى المؤسسات الحكومية او الخاصة وجلب الاجهزة والادوات لتدريب الطلبة عليها في الكليات التطبيقية.
- التوسع في نطاق التدريب العملي وتحويل الدراسة من الدراسة النظرية البحتة للتطبيق العملي.

(٢)



سلطنة عمان
وزارة التعليم العالي
كلية العلوم التطبيقية بصور

(نناشد بالعدالة... فحقوقنا لمسنا ... فكن معنا)

- يجب إختيار هيئة أكاديمية ذات كفاءة تدريسية عالية .
- يجب مراقبة الهيئة الأكاديمية من ناحية التفاضل بين الطلاب والتمييز بينهم .
- إلغاء نظام المنحنى وتوزيع الدرجات بصوره واضحة ومراقبة من قبل الادارة والسماح بوجود الطلاب أثناء مراجعة أوراق التظلم .
- تحديد الوقت للطلاب أثناء الإختبارات القصيرة حتى يتسنى لهم التأكد من إجاباتهم على الأقل .
- المساواة في الخدمات الطلابية بين كليات العلوم التطبيقية وجامعة السلطان قابوس.
- تطوير مختبرات الحاسب الآلي ونظام شبكة (الوابر لس) وزيادة عدد الحواسيب والأجهزة المرفقة بها.
- الأخذ بعين الأعتبار بالتقييم الذي يقوم به الطلاب للمدرسين على موقع بلاك بورد.
- الإرتقاء بالخدمات الإدارية في كل من (الإدارة ، المالية، القبول والتسجيل، والهيئة الاكاديمية، ومركز الخدمات الطلابية).
- منح شهادات التخصص الفرعي (المايز) بالمواد المدروسة.
- إنشاء مظلات وزيادة المعشب مع توفير الكراسي والطاولات للطلبة والطالبات .
- وضع جداول إمتحانات تناسب مع مقررات الطلاب الدراسية.
- أن يكون حفل التخرج في مكان واحد مفتوح لجميع التطبيقيين وبث الحفل مباشرة عن طريق التلفاز .
- توفير وجبة غداء مجانية لكل الطلاب .
- الإهتمام بنظافة المرافق العامة في الكلية وخاصة دورات المياه .
- إعادة تهيئة المباني القديمة وتوفير سكنات داخلية للطالبات .
- زيادة الرواتب الى 120 ريال مساوة بإخوانهم في مؤسسات التعليم العالي .
- إنشاء مركز ترفيهي لطلاب والطالبات .
- تطوير المطعم ومرفقاته من أطعمة وغيرها مع توسعته .
- مراعاة طالبات السكنات الداخلية وتنظيم رحلات إسبوعية للترفيه عن النفس .

... نرجوا النظر إلى مطالبنا بعين الإعتبار حتى ندفع عجلة تطور كليتنا الى الامام ...

B+S

A translation of the highlighted points in the flyer above
The flyer was addressed to "whom it may concern" and entitled "The Colleges of Applied Sciences-Sur Students' Requests". The highlighted points concerning assessment processes are translated below: 1- Disallowing teachers to be subjective in marking or be influenced by any personal issues with the students. 2- Changing the producers followed in calculating the GPA; so a grade 2.3 is designated a B grade instead of a C+. 3- When a student appeals for a reappraisal, his test paper should be reviewed by a special committee in which the course instructor is a member. 4- Students with a GPA under probation should be given more than three semesters to elevate their GPA. All students who were dismissed because their GPA was 1.5 or less should be allowed to resume study. A GPA of 1.8 should be the least value qualifying for an "under probation" status instead of the current 2.0. 5- Eliminating norm-referenced assessment and awarding grades based on student's achievements. 6- Extending time allocated for English language examination sessions to allow students to revise their responses to the test tasks. 7- In setting exam timetables, the amount of materials students have to study for each test should be considered and students should be given enough time to study.

Appendix. 4.1. A leaflet about the phases and aims of the study distributed to students and teachers

English Language Assessment in a Higher Education Institutions in Oman: A case study

Introduction

There is a considerable debate about the role of English language proficiency in second/foreign language students' academic achievement amongst teachers, academics, and policy makers. The situation is similar in Oman in which English language is the medium of instruction in almost all higher educational institutions. How much English language proficiency influence academic achievement and how should the English language be assessed in away that informs decision makers about the right levels of students' competences to handle academic studies with out being hindered by linguistic barriers. Who the stakeholders are and what roles they play are all important factors in studying language assessment.

I am a PhD student at the University of Edinburgh and would like to understand how English language assessment functions in Omani higher education institutions. Your participation will contribute immensely in shaping and constructing my study.

Context

This study intends to include participants from the foundation Program and First Year (FY) students at Sur and Rustaq Colleges of Applied Sciences in Oman. In CAS, students spend almost one academic year to attain the required level of English language proficiency before being admitted to FY. I plan to follow students in their last semester in FP and first semester in FY to obtain a more comprehensive understanding of the English language assessment and students experiences. I will also be in contact with the FP and FY English and academic teachers.

Participants

Phase1 (Feb2011-May 2011)

- Students

FP level A students who agree to participate in the second phase

- Teachers

FP teachers (not necessary to participate in phase two)

Phase2 (Sept 2011-Dec.2011)

- Students

FY students who participated in phase 1

-Teachers-

FY English language Teachers -

& FY academic courses teachers-

How will you participate?

1- A Questionnaire

Students in Arabic (10-15 minutes)

Teachers (10-15 minutes)

2- Interview/ Focus interview

Teachers will be interviewed individually. (15-30 minutes)

Students will be interviewed in focus groups consisting of maximum 15 students in Arabic (45-60 minutes)

Confidentiality and Security

The data collected in this study will be accessed by the researcher only and will be used for the purposes of this study. Confidentiality and anonymity are promised to the participants.

Please contact me for futher information on:

Fatimaalhajri@gmail.com

Fatma Al Hajri (PhD student)

Appendix 4.2. Informed consent distributed to participants prior to conducting the study

English Language Assessment in the Colleges of Applied Sciences (CAS) in Oman: A mixed methods case study

Information consent for participants:

Introduction:
You are invited to participate in a study about language assessment predictability in entry and EAP exams. This study will be conducted by Fatma Al Hajri, doctoral student in Moray House School of Education, University of Edinburgh. The results of the study will contribute to Fatma's dissertation in partial fulfilment a doctorate in language assessment. Your participation in this study is voluntary.

Purpose:
In this study, I intend to evaluate the English language assessment in the foundation year (FP) and the first year (FY). I hope that this study will highlight the validity of language assessment and its correlation with students' performance in the first year content courses namely; communication, IT and IBA.

Procedures:
Consenting to participate in this study entails responding to a questionnaire and being interviewed for approximately 30 minutes by Fatma Al Hajri. Your comments on the questionnaire and interview will be highly considered and will participate in understanding and evaluating the assessment process.

Voluntary participation:
Participation in this case study in entirely voluntary. You may refuse to participate or withdraw at any time.

Confidentiality:
All data collected through questionnaires and interviews will be highly confidentially and will not be used in any way but for the purposes of the study. Also, all participants will be anonymous and coded using numbers or pseudonyms when referred to in reporting and analysing the data.

Contact:
If you have any comments, please contact me on:
Mobile: 0096892829100
E-mail: Fatimaalhajri@gmail.com

Appendix 4.3. Students' Questionnaire Topics and Items in Phase 1

Topics	Sub-topic	Number of Questions	Items
Perceived validity	Content	4	. There is a strong connection between what we do and learn in the classroom and the final test. . There is a strong connection between what we do and learn in the classroom and the continuous assessment. . I understand how my language performance will be assessed on the FP. . The assessment instruments provide me with sufficient feedback on my English language performance.
	Construct/ General	1	. My scores in language assessment reflect my real achievement level on the FP.
	Construct/ Test	2	. Tests administered during the FP have assisted me to function in English in real life. . Tests administered during the FP have assisted me to function in English in my academic studies.
	Construct/ Continuous Assessment	2	2.4. Continuous assessment on the FP has assisted me to function in English in real life. 2.5. Continuous assessment on the FP has assisted me to function in English in my academic studies.
Preference for tests		2	. I would prefer to have a final test only instead of continuous assessment and a final test. . Some sections of the continuous assessment should be changed.
Preference for continuous assessment		2	. Continuous assessment provides me with a better opportunity to demonstrate my English language skills than the tests. . Some sections of the final test should be changed.
Satisfaction with current assessment tools		3	. I am satisfied with the assessment instruments used to evaluate my English language skills. . FP assessment instruments should not have fewer assessment parts (tests, presentation, written report, quizzes ...etc.). . The assessment instruments should be changed to include aspects of students'

			English language that are not assessed currently. (Recode)[21]
		2	. The division of scores assigned to the different Foundation Programme assessment instruments (i.e. tests, quizzes, reports and presentations) is appropriate. 6.2. Usually the difference between my scores in the tests and continuous assessment is not considerable.
Impact	Social (positive)	5	. Tests on the FP make me feel stressed. (Recode) . Continuous Assessment on the FP make me feel stressed. (Recode) . English language assessment on the FP is fair. . English language in the FP assessment is not frightening to me. . Passing the FP assessment does not depend on luck or supernatural powers.
	Political	2	. Being taught and assessed in English creates more employment opportunities for me. . Being taught and assessed in English makes Oman an active part of the global village.
Total		**25**	

Appendix 4.4. Student Questionnaire Topics and Items in Phase 2

Topic	Number of Questions	Items
Dissatisfaction with FP assessment	3	. Assessment on the FP should have allowed more students to proceed to the FY. . Assessment instruments should be changed on the FP to better match my English language needs in academic courses. . Assessment instruments should be changed in the FY to better match my English language needs on academic courses.
Adequacy of English language level for First Year study	4	. My English language level is adequate to understand the academic courses and to meet their assessment requirements. . I have difficulty understanding my lecturers in the FY academic courses because my English language level is insufficient. (Recoded) . I have difficulty in expressing my ideas in writing in the academic course

[21] "Recode" signifies that the responses to the item were recoded when entered in SPSS to conform to the other items in terms of the general meaning.

		assessments. (Recoded) . I have a difficulty in understanding the reading passages for the academic courses assessments. (Recoded)
Predictive validity	2	. The better a students' English language ability, the better his/her achievement in academic courses will be. . I needed more English language courses if I am to perform well in the First Year.
First Year assessment Construct validity	2	. Assessment instruments in the FY measured my language skills appropriately. . In the English language course, teachers assess both my ideas and my language.
Consequence and impact	3	. The assessment and teaching in English creates more employment opportunities for me. . Teaching and assessing in English at university level supports my country's status internationally. . FP assessment has more negative social consequences to me than FY assessment.
Assessing English language and ideas in academic courses	7	. Teachers on academic courses should assess students on their written expressions as well as their ideas. . I would like to get feedback on both my ideas and my written expression in academic courses. . In academic courses, students should not be marked for their English language skills. (Recoded) . Academic course teachers assess both my ideas and my language. . I think that assessment in the academic courses should not require written assignments in English.
Assessing English language and ideas in English language courses	2	. In the English language course, teachers assess both my ideas and my language. . I would like to get feedback on both my ideas and written expression in English language courses.

Appendix 4.5. Teacher Questionnaire's Topics and Items in Phase 1

Topic	Sub-topic	Number of Questions	Items
Reliability		3	. The criteria and the rating scales that students are assessed by on the FP facilitate

			assessing students consistently. . The assessment instruments on the FP are consistent in evaluating students' language performance. . I am satisfied with the reliability of the assessment instruments implemented on the FP.
Validity	Content	3	. The scores on the different assessment instruments reflect the time spent on teaching the English language skills. . The assessment instruments in the FP represent the English language skills and activities covered in the curriculum appropriately. . The assessment instruments efficiently represent the FP objectives.
	Predictive	2	. The FP assessment instruments report on student's abilities to linguistically handle FY academic courses. . The FP English assessment prepares students well to cope with the language demands of their academic courses.
	Face (negative)	3	. The assessment instruments used on my courses are appropriate in assessing students' English language abilities. (Recode) . The FP assessment instruments should be changed. . There should be fewer assessment instruments (continuous assessment and tests) in FP courses than there are currently.
	Construct	3	. The current assessment instruments are valid. 0. The FP assessment instruments provide teachers with suitable information about their students' English language performance. 1. The students' scores on the FP assessment represent their language performance levels accurately.
Test / continuous assessment		2	. Tests are more valid than continuous assessment. . Tests are more reliable than continuous assessment.
Preference of centrality in writing assessment		3	. Teachers should write their own final tests locally at the colleges. (Recode) . Teachers should undertake their own continuous assessment locally at the colleges. (Recode) . Teachers should conduct the same assessment instruments in all of the six colleges.
Confidence in marking and writing assessment		3	. I am confident of my ability to write final tests and assessment activities for FP students.

			. I need more training to write reliable and valid assessment instruments. (Recode) . I have appropriate experience in marking tests and assessment tasks using provided scales.
Impact	Social	5	. I have made the students aware of the consequence of failing/passing FP assessment. . As far as I know, the department is taking sufficient account of the probable social consequences of failure in the FP assessment of students (e.g. making students and teachers aware of those consequences, working to avoid the severity of the consequences). . The assessment instruments are fair to students and should be carried out in the same way in the future. . I have the opportunity to give feedback on the quality of the assessment instruments. . Other parties (students, society, researchers and other organizations) have the opportunity to give feedback on the quality of the assessment activities and tests.
	Political	3	. The Omani National Standards for the FP and the FP audit are vital to ensure accountability in English language teaching institutions. . Assessing the academic courses in English helps to develop the country's economy. . I think students' scores in FP English language assessment should not be used as a gate-keeper to higher education in Oman. (Recode)
Total		**30**	

Appendix 4.6. Teacher Questionnaire Topics in Phase 2

Topic	Sub-topic	Number	Items
Consistency between First Year and Foundation Program English Language Assessment		4	. In general, assessment of student performance on the FP and in the FY English course provides similar results for the majority of students. . There is a close relationship between student performance on the FP and in their performance in the FY English course. . Assessment should be standardised within CAS. . There is a close correlation between student performance on the FP and their performance in FY academic courses.
			Students do better in FY academic courses when their English language scores on

Foundation Program Assessment Validity	Predictive	5	the FP are higher. If students perform well in English language courses, they will perform well in First Year academic courses too. Students' weak performance in FY academic courses could be caused by factors other than their English language levels. (Recode) The low English language level of some students in the FY causes them to fail FY academic courses. Students' language levels influence their achievement in FY academic courses.
	Construct	2	When allowing students to pass into the F Y, it is more informative to focus on students' results in individual English language skills on the FP (e.g. writing, listening or reading marks) than on their total marks. Students' scores in all language skills (reading, writing, speaking and listening) assessment on the FP are equally important indicators of their future academic achievement in the FY.
Satisfaction with assessment	FP	3	Most students admitted to the FY have the appropriate English language skills to understand and communicate in their academic courses. Students' current language abilities are generally adequate for the academic courses in the FY. English language assessment on the FP effectively measured students' abilities to function in FY academic courses.
	FY	2	Assessment instruments in the FY English course focus on the academic language skills students need in FY academic courses. FY English course assessment measures students' academic language use efficiently.
Assessing English language in the First Year academic courses		5	Assessment criteria in the academic courses should not include students' English language level. (Recoded) One of the criteria used to mark the FY academic courses should be English language competence. Academic course assessment should aim to be less dependent on students' language ability.(Recoded) When assessing academic courses, markers should overlook language inaccuracies as long as the meaning is clear.(Recoded) Teachers in academic courses should assess students on their written expressions

			as well as their ideas.
Social impact		3	. The current assessment instruments take account of other parties' opinions (e.g. students). (Recode) . Planning how to assess students' work is a process to which teachers, students, society and other related organizations should contribute. (Recode) . In my department, students' opinions about assessment instruments are considered in the design of assessment instruments.(Recode)
Political impact		3	. English language assessment should not be a gate-keeper to higher education in Oman. (Recode) . Assessing and teaching in English creates more employment opportunities for students. . Being taught and assessed in English makes Oman an active part of the global village.
Total		**27**	

Appendix 4.7. A Sample of the Questionnaires used

A. Student Questionnaire in Phase 1 in English

A questionnaire about English language assessment at the Colleges of Applied Sciences Foundation Programme students (Phase 1)

Fatma Al Hajri (a PhD candidate at the University of Edinburgh)

Thank you for participating in this study. This questionnaire is about English language assessment instruments in the *Foundation Program* **(FP)**. Would you please help me understand how students language abilities are evaluated using continuous assessment and a final test by responding to the questionnaire; all of the information collected by this questionnaire will remain confidential. Participation is voluntary at all times.

The questionnaire includes **three sections**:

(1) Participant's information
(2) Assessment instruments
(3) General views about assessment instruments

For the purpose of this study, the words **test, continuous assessment and assessment instruments** will be used as follows:

Tests	The final test and quizzes
Continuous assessment	Written project, presentations, classroom activities, and projects.
Assessment instruments	Both tests and continuous assessment.
FP	Foundation Program – English language courses only
FY	First Year

Section (1): Please circle the most appropriate answer.

Gender	(1) Female		(2) Male		
ID:……..					
Age	(1) 18	(2) 19	(3) 20	(4) 21	(5) 22
Specialization	(1) IT	(2) IBA	(3) Design	(4) Communication	(5) English Language
English Language level	(1)Poor	(2)Fair	(3)Good	(4)Very good	(5)Excellent

Section (2): Please circle the most appropriate answer to give your honest opinion. There is no a right or wrong answer.

Statements	***strongly agree***	***agree***	***no opinion***	***disagree***	***strongly disagree***
1- I am satisfied about the types of assessment instruments used to evaluate my English language skills.	***1***	***2***	***3***	***4***	***5***
2- I would prefer to have a final test only	***1***	***2***	***3***	***4***	***5***

instead of continuous assessment and a final test.					
3- The continuous assessment provides me a better opportunity to show my English language skills compared to the tests.	***1***	***2***	***3***	***4***	***5***
4- Some sections of the final test should be changed.	***1***	***2***	***3***	***4***	***5***
5- Some sections of the continuous assessment should be changed.	***1***	***2***	***3***	***4***	***5***
6- The marks awarded to the different FP assessment instruments like the tests, quizzes, reports and presentations are appropriate.	***1***	***2***	***3***	***4***	***5***
7- There is a strong connection between what we do and learn in classroom and the final test.	***1***	***2***	***3***	***4***	***5***
8- There is a strong connection between what we do and learn in classroom and the continuous assessment.	***1***	***2***	***3***	***4***	***5***
9- Tests in the FP make me feel stressed.	***1***	***2***	***3***	***4***	***5***
10- CA in the FP make me feel stressed.					
11- FP assessment instruments should not have fewer different parts (tests, presentation, written report, quizzes ...etc.).	***1***	***2***	***3***	***4***	***5***
12- Usually the difference between my scores in the tests and continuous assessment is not considerable.	***1***	***2***	***3***	***4***	***5***
13- I understand how my language performance will be assessed in FP.	***1***	***2***	***3***	***4***	***5***
14- The assessment instruments should be changed to include aspects of students' English language that are not assessed currently.	***1***	***2***	***3***	***4***	***5***
15- English language assessment in FP is fair.	***1***	***2***	***3***	***4***	***5***
16- English language in FP assessment is not frightening to me.	***1***	***2***	***3***	***4***	***5***
17- Passing the FP assessment does not depend on luck or supernatural powers.	***1***	***2***	***3***	***4***	***5***
18- The assessment instruments provide me with enough feedback on my English language performance.	***1***	***2***	***3***	***4***	***5***
19- Being taught and assessed in English creates more employment opportunities for me.	***1***	***2***	***3***	***4***	***5***
20- Being taught and assessed in English makes Oman an active part of the global village.	***1***	***2***	***3***	***4***	***5***
21- My scores in language assessment reflect my real achievement level in FP.	***1***	***2***	***3***	***4***	***5***
22- Tests in FP assist me to function in English in real life.	***1***	***2***	***3***	***4***	***5***
23- Continuous assessment in FP assists	***1***	***2***	***3***	***4***	***5***

me to function in English in real life.					
24- Tests in FP assist me to function in English in academic studies.	***1***	***2***	***3***	***4***	***5***
25- Continuous assessment in FP assist me to function in English in academic studies	***1***	***2***	***3***	***4***	***5***

Section (3): I will be grateful if you could help me reach a better understanding of the following issues.

1- Do you think that FP assessment instruments should be changed in anyway? If possible please elaborate in your answer (20-50 words)

2- I would be grateful if you add any other comment that you think is relevant to the issues raised.

Thank you for participating in this questionnaire, to contact the research please e-mail to: fatimaalhajri@gmail.com

B. Student Questionnaire phase 1 (the Arabic language version)

استبانه عن كيفية تقييم مستوى اللغة الإنجليزية في السنة التأسيسية في كليات العلوم التطبيقية للمستوى(أ)-Level A

الاستبانه جزء من دراسة الدكتوراه في جامعة أدنبره- المملكة المتحدة

للباحثة: فاطمة بنت سعيد الحجري

أتقدم بداية بالشكر الجزيل لك أخي الكريم / أختي الكريمة على تخصيص جزء من وقتكم لتعبئة هذه الاستبانه.

يتعرض موضوع البحث الى دراسة أدوات التقييم مثل الإمتحانات النهائية و القصيرة و البحوث و المشاريع و العروض الكلامية المستخدمة لقياس المستوى التحصيلي في مقرري اللغة الإنجليزية لطلبة السنة التأسيسية في كليات العلوم التطبيقية، وسوف يتم استخدام البيانات بسرية تامة لغرض الدراسة فقط علما بأن أسماء المشاركين سيعبر عنها بأرقام أو اسماء رمزيه عند مناقشة نتائج الدراسة. كما ان المشاركة في هذه الدراسة تعتبر تطوعية ويمكن للمشاركين الإنسحاب منها في أي وقت يرغبون بذلك. يرجى الإجابة عن جميع الأسئلة ليتسنى استخدام المعلومات بدقة.

تنقسم هذه الإستبانه إلى ثلاثة أجزاء:

الجزء الأول: أسئلة عن الشخص المشارك في الإستبيانه
الجزء الثاني: أسئلة إختيار من متعدد عن أدوات التقييم
الجزء الثالث: أسئلة مفتوحةعن أدوات التقييم

الجزء الأول: أجب على التالي بما هو مناسب.

1- **الرقم الجامعي**:_____

2- الجنس:
1- ذكر 2- أنثى

3- العمر
1- 17 2- 18 3- 19 4- 20

4- التخصص:
1- تكنلوجيا المعلومات 2- تصميم 3-دراسات الإتصال 4- إدارة الأعمال الدولية

5- مستوى اللغه الإنجليزية (تبعا للتقيمك لنفسك)
1- ضعيف 2- مقبول 3-جيد 4-جيد جدا 5- ممتاز

الجزء الثاني: إختر إستجابة واحده لكل من العبارات التالية بما يتوافق و إعتقادك

العبارات	موافق بشدة	موافق	محايد	معارض	معارض بشدة
1- يتم تقييم مهاراتي في اللغة الإنجليزية بشكل **جيد** بواسطة أدوات (إمتحانات منتصف ونهائية وبحوث ومشاريع) التقييم المستخدمة حاليا.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
2- التقييم بواسطة الامتحان النهائي أفضل من التقييم المستمر.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
3- يوفر لي التقييم المستمر(مثل البحوث و المشاريع و العروض الكلامية... الخ) فرصة عرض و توضيح مهاراتي في اللغة الإنجليزية أكثر مما يوفره الإمتحان النهائي.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
4- يجب تغييربعض الأجزاء من الامتحان النهائي .	موافق بشدة	موافق	محايد	معارض	معارض بشدة
5- يجب تغيير بعض الأجزاء من التقييم المستمر (مثل البحوث و المشاريع و العروض الكلامية... الخ).	موافق بشدة	موافق	محايد	معارض	معارض بشدة
6- توزيع الدرجات الحالي على أدوات التقييم (الإمتحانات و التقييم المستمر) مناسب بالنسبة لي.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
7- توجد علاقة قوية بين ما نتعلمه و نفعله في المحاضرة من مهارات لغوية و بين المهارات المتضمنة في الامتحان النهائي	موافق بشدة	موافق	محايد	معارض	معارض بشدة
8- توجد علاقة قوية بين ما نتعلمه و نفعله في المحاضرة من مهارات لغوية و بين المهارات المتضمنة في التقيم المستمر (مثل البحوث و المشاريع و العروض الكلامية... الخ).	موافق بشدة	موافق	محايد	معارض	معارض بشدة
9- أداء الإمتحانات يصيبني بالتوتر.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
10- أداء التقييم المستمر (المستمر (مثل البحوث و المشاريع و العروض الكلامية... الخ) يصيبني بالتوتر.					
11- يجب أن لا يقل عدد الامتحانات و البحوث والعروض الكلامية وغيرها المستخدمة حاليا في السنة التأسيسية.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
12- لا يكون هناك فرق بين درجاتي في الامتحانات و درجاتي في البحوث و المشاريع وغيرها.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
13- أنا أعرف كيف سيتم تقييم مهاراتي في اللغة الإنجليزيه في السنة التأسيسية.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
14- يجب تغيير أدوات التقييم (من إمتحانات و بحوث و مشاريع) لتضم بعض المهارات اللغوية التي لايتم تقييمها حاليا.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
15- التقييم في السنة التأسيسية عادل.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
16- التقييم في السنة التأسيسية لا يثير الرعب.	موافق بشدة	موافق	محايد	معارض	معارض بشدة

العبارات	موافق بشدة	موافق	محايد	معارض	معارض بشدة
17- لا يعتمد النجاح في السنة التأسيسة على الحظ.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
18- توفر لي الأدوات التقييم (الامتحانات و البحوث و المشاريع) معلومات كافية عن المستوى اللغوي الخاص بي.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
19- أداء الإمتحانات و البحوث و المشاريع باللغة الإنجليزية سيخلق لي المزيد من الفرص للحصول على وظيفة مستقبلا.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
20- أداء الإمتحانات و البحوث و المشاريع باللغة الإنجليزية في المرحلة الجامعية يعزز دور السلطنة في المجتمع الدولي.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
21- تمثل درجاتي وتحصيلي في أدوات التقييم المختلفة صورة مقاربة لتطور مهارتي في اللغة الإنجليزية في السنة التأسيسية.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
22- تساعدني الإمتحانات على إستخدام اللغة الإنجليزيه في الحياة اليومية بصورة عامة.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
23- تساعدني البحوث و المشاريع على إستخدام اللغة الإنجليزيه في الحياة اليومية بصورة عامة.	موافق بشدة	موافق	محايد	معارض	معارض بشدة
24- تساعدني الإمتحانات على إستخدام اللغة الإنجليزيه في الحياة الأكاديمية (السنة القادمة).	موافق بشدة	موافق	محايد	معارض	معارض بشدة
25- تساعدني البحوث و المشاريع على إستخدام اللغة الإنجليزيه في الحياة الأكاديمية (السنة القادمة).	موافق بشدة	موافق	محايد	معارض	معارض بشدة

الجزء الثالث: الرجاء الإجابة على الأسئلة بما تراه مناسبا

1- هل تعتقد بوجوب تغيير ادوات التقييم في السنة التأسيسية للغة الإنجليزية؟ هلا تكرمت بشرح إجابتك؟

2- إذا كان لديك ماتضيفيه عن هذا الموضوع فتفضل بإضافتة في الفراغ مشكورا.

Appendix 4.8. Researcher's Responses to a Research Ethics Checklist from the College of Humanities and Social Sciences, University of Edinburgh

Research ethics checklist

2 RISKS TO, AND SAFETY OF, RESEARCHERS	
Those named above need appropriate training to enable them to conduct the proposed research safely and in accordance with the ethical principles set out by the College	Yes/No√
Researchers are likely to be sent or go to any areas where their safety may be compromised	Yes/No√
Could researchers have any conflicts of interest?	Yes/No√
3 RISKS TO, AND SAFETY OF, PARTICIPANTS **Could the research induce any psychological stress or discomfort?**	Yes√/No It might be considered by some as an act of evaluating their performance. The researcher will introduce the participants to the study in advance.
Does the research involve any physically invasive or potentially physically harmful procedures?	Yes/No√
Could this research adversely affect participants in any other way?	Yes/No√
4 DATA PROTECTION	
Will any part of the research involve audio, film or video recording of individuals?	Yes√/No Consent will be attained.
Will the research require collection of personal information from any persons without their direct consent?	Yes/No√
How will the confidentiality of data, including the identity of participants (whether specifically recruited for the research or not) be ensured?	As the researcher will be the sole person to use the data, confidentiality will be ensured and identities will never be exposed in any way. Also, the data will be stored in a safe place.
Who will be entitled to have access to the raw data?	Only the researcher
How and where will the data be stored, in what format, and for how long?	The data will be stored in the researcher university office and will be locked in a locker
What steps have been taken to ensure that only entitled	The researcher intends to do all

persons will have access to the data?	data analysis by herself and will keep the data in a secure location accessed by her only.
How will the data be disposed of?	Once the research has finished her PhD thesis and any other intended publication on the topic in the following five years, the data will be disposed by shredding all paper based data and damaging the soft data.
How will the results of the research be used?	The results will be used to add to the current understanding of the literature on assessment. They will be also used to engender recommendations for the ministry of higher education about language assessment.
What feedback of findings will be given to participants?	The participants will have access to the final thesis as it will be kept in the Ministry of Higher education archive
Is any information likely to be passed on to external companies or organizations in the course of the research?	Yes/No√
Will the project involve the transfer of personal data to countries outside the European Economic Area?	Yes/No√
5 RESEARCH DESIGN	
The research involves living human subjects specifically recruited for this research project If 'no', go to section 6	Yes√/No
How many participants will be involved in the study?	Teachers:50-70 Students:150
What criteria will be used in deciding on inclusion/exclusion of participants? How will the sample be recruited?	The topic of the study decided on the features of the possible study candidates. The researcher will include those how are willing to participate unless the number exceeds the expectation. If it does, the sample will be recruited randomly
Will the study involve groups or individuals who are in custody or care, such as students at school, self-help groups, residents of nursing home?	Yes/No√
Will there be a control group?	Yes/No√
What information will be provided to participants prior	Information Leaflet will be

to their consent? (e.g. information leaflet, briefing session)	distributed to participants and a consent form
Participants have a right to withdraw from the study at any time. Please tick to confirm that participants will be advised of their rights.	√
Will it be necessary for participants to take part in the study without their knowledge and consent? (e.g. covert observation of people in non-public places)	Yes/No√
Where consent is obtained, what steps will be taken to ensure that a written record is maintained?	Participants will sign a form
In the case of participants whose first language is not English, what arrangements are being made to ensure informed consent?	The informed consents are translated into Arabic as well as the information leaflet and questionnaires for the students
Will participants receive any financial or other benefit from their participation?	Yes/No√
Are any of the participants likely to be particularly vulnerable, such as elderly or disabled people, adults with incapacity, your own students, members of ethnic minorities, or in a professional or client relationship with the researcher?	Yes/No√
Will any of the participants be under 16 years of age?	Yes/No√
Do the researchers named above need to be cleared through the Disclosure/Enhanced Disclosure procedures?	Yes/No√
Will any of the participants be interviewed in situations which will compromise their ability to give informed consent, such as in prison, residential care, or the care of the local authority?	Yes/No√
6 EXTERNAL PROFESSIONAL BODIES	
Is the research proposal subject to scrutiny by any external body concerned with ethical approval? If so, which body?	Yes/No√

Appendix 5.1. List of Documents Analysed in Both the First and Second Phases

First Phase

1	Essay Plan Example (Foundation Programme)
2	Research Log (1) (Foundation Programme)
3	Research Log (2) (Foundation Programme)
4	Essay Plan (Foundation Programme)
5	How to prepare a title page (Foundation Programme)
6	Reference list (Foundation Programme)
7	Presentation guidelines (Foundation Programme)
8	Presentation preparation sheet (Foundation Programme)
9	2010-SUM-E6001-SB-V1-combined SB&AB*
10	2011- SPR- MT- E6001-SB- V1 + V2- D1*
11	2011-SP-E6001-SB-V2- D2 –FINALS*
12	2011-SPR-AK-V2-D2*
13	2011-SPR-E6001-AK-V1 -D2*
14	2011-SPR-E6001-V1-V2-D2*
15	2011-SPR-MT-E6001 SB -V1-D1*
16	2011-SPR-MT-E6001-AK-V1-D1*
17	2011-SPR-MT-E6001-AK-V2-D1*
18	2011-SPR-MT-E6001-SB-V2*[i]
19	Academic Calendar Spring 2011.
20	Colleges of Applied Sciences: Academic Regulations (Arabic)
21	AES project presentation criteria
22	Analysis of MTs* results
23	Assessment planning schedule
24	AUT-2010-IBRI 20Item Analysis*
25	CAS English Department Assessment Handbook V2
26	Audit Report for CAS-Ibri
27	Audit Report for CAS-Salalah
28	Audit Report for CAS-Sohar
29	Course Specifications for Foundation English
30	Students' Scores in the Last Year of High School
31	Draft Marking Scale for AES report
32	ENGL 6001 - Specifications
33	FA-SPRING11-SUR(1)*
34	FN A April 22nd 2011-edit*
35	FN A Master*
36	FN A Rustaq Mid
37	FN_A_Results Sur edit 1
38	FN_A_Results, Sur edit 2
39	FN_A_ResultsSur edit 3
40	FN_A_Results Rustaq
41	Foundation A Final Exam Writing Rating Scale 06 10
42	Foundation Year AS project Topics - level A
43	Foundation Year Academic Skills Project Outline

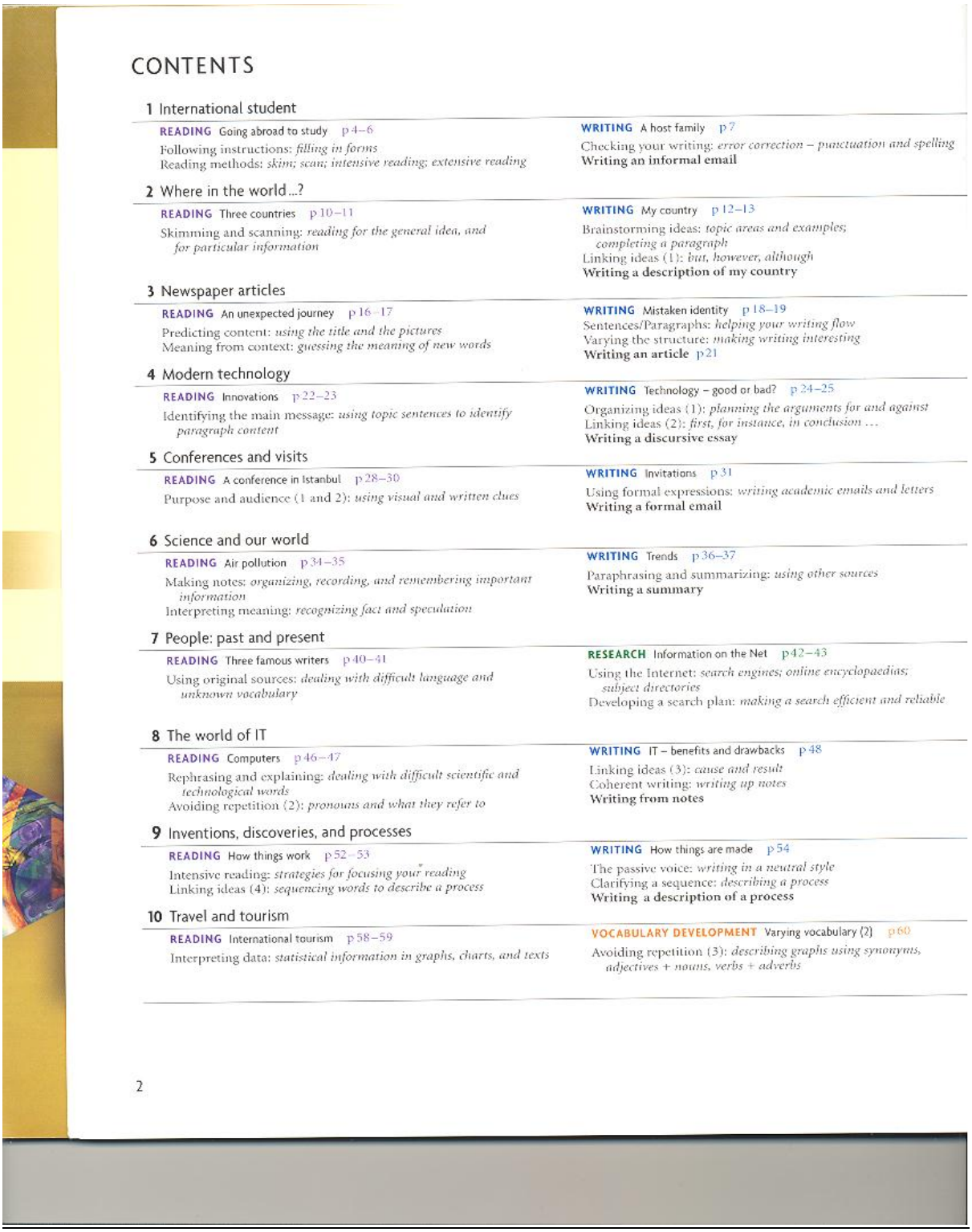
44	Foundation Program: 2010-11
45	Foundation Program Calendar (Winter)
46	Oman Academic Standards for General Foundation Programs
47	Employed CAS Graduates in 2011
48	Minutes - joint Assessment Team and Foundation Team 14-5-11
49	Placement Test
50	English Department Anti-Plagiarism Procedures: Student plagiarism V3, 02/11
51	Academic Regulations- English
52	Speaking Test Assessment Criteria Foundation
53	Speaking Test Questions A Part 3
54	Speaking Test Questions A, B part 2
55	Speaking Test Questions A,B,C Part 1
56	Speaking test score sheet A
57	Speaking test rubric A
58	Specs for Spread sheets Foundation
59	SPR-2011-IBRI Item Analysis*
60	Student Guide _English
61	The New Foundation Program
62	Enrolled Students in CAS 2010-2011
63	CAS Statistics Booklet 2011
64	Report form AES A
65	Report form GES A
66	Rustaq and Sur Students' Civil Numbers
67	2011-SPR-4001-AK-V2-D1-MT
68	2011-SPR-4001-AK-V2-D1-MT
69	2011-SPR-4001-SC-Mid-term[1]
70	2011-SPR-MT-E5001-LS-V1-D3[1]
71	2011-SPR-MT-E6001 SB -V1-D1[2]
72	2011-SPR-MT-E6001-V2
73	SPR-2011-5001-AK-V2-MT*
74	2011- SPR- MT- E6001-LS- V1- D1@

Second Phase

75	Oral Presentation Assessment Criteria for ENGL1111
76	ENGL1111 Purpose Statement
77	Assessment Policies Procedures- English Department
78	Course Specification for ENGL 1111 AUT 2011
79	ENGL 1111 Project booklet
80	ENGL 1111 Project Specifications
81	ENGL1111Specifications
82	Final Exam BUSN1400-2009-version2
83	First Year Vocabulary List

84	Grammar Review Portfolio
85	IT Final Exam in 2009
86	New Academic Calendar ENGL1111
87	Sample Topics and Questions for ENGL 1111 & ENGL 1222 Project
88	Specifications for Spread sheet ENGL1111 & 1222
89	ENGL1111 Project Report Rating Scale
90	Course Outline for Fundamentals of IT (INFT1001)
91	Communication Interview Questions
92	Exam Marking Guide
93	Exam Additional Hand-out (Communications)
94	Oman Communication Exam
95	Oman Communication Handbook
96	Informal Peer Feedback Sheets
97	Informing Talk Preparation
98	Persuasive Talk Preparation
99	Assertiveness Scenarios
100	Becoming Assertive
101	Cross-cultural Communication Barriers Hand-out
102	Lecturer Feedback Form
103	Nasa Exercise
104	Oman Consultancy Report
105	Student Guide Arabic
106	Policies on Plagiarism
107	Exam Instructions
108	Student Guide Arabic
109	Intake of Secondary Graduate in University and Colleges.
110	Colleges of Applied Sciences: Academic Regulations (English)
111	Headway Academic Skills (Level 2)
112	Headway Plus (Intermediate)
113	CAS English Department Assessment Handbook
114	Assessment Policies: English Department October 2011
115	Business Fundamentals Course handbook
116	Business Fundamentals: Final Exam
117	Course Outline for Fundamentals of IT (INFT1001)
118	IT Final Exam

Appendix 5.2. The Contents of the Headway Academic Skills (level 2) textbook, used in the AES course in the Foundation Programme



CONTENTS

Appendix 5.3. The Contents pages of the Headway Plus Intermediate Textbook, used in the GES Course of the Foundation Programme

CONTENTS

LANGUAGE INPUT

UNIT	GRAMMAR	VOCABULARY	EVERYDAY ENGLISH
1 It's a wonderful world! p6	**Auxiliary verbs** *do, be, have* p7 **Naming the tenses** Present, Past, Present Perfect p7 **Questions and negatives** *What did you do last night?* *Cows don't eat meat.* p7 **Short answers** *Yes, I did.* p8	What's in a word? Parts of speech and meaning Spelling and pronunciation Word formation Words that go together Keeping vocabulary records p12	Social expressions *Never mind.* *Take care!* *You must be joking!* p13
2 Get happy! p14	**Present tenses** Present Simple *Does she work in a bank?* p15 Present Continuous *Is he working in France at the moment?* p15 Simple or continuous? *She usually drives to work, but today she isn't driving. She's walking.* p17 Present passive *We are paid with the money people give.* *Children are being treated with a new kind of medicine.* p18	Sport and leisure *play football* *go sailing* *do aerobics* p20	Numbers and dates *Money, fractions, decimals, percentages, dates, phone numbers* p21
3 Telling tales p22	**Past tenses** Past Simple and Continuous *He danced and sang.* *He was laughing when he saw the baby.* p23 Past Simple and Past Perfect *I didn't laugh at his joke.* *Why? Had you heard it before?* p24 Past Passive A Farewell to Arms *was written by Ernest Hemingway.* p27	Art and literature *painter* *poet* p25 Collocations *paint a picture* *read a poem* p25	Giving opinions *What did you think of the play?* *It was really boring! I fell asleep during the first act.* p29
Stop and Check 1 Teacher's Book			
4 Doing the right thing p30	**Modal verbs (1) – obligation and permission** *have (got) to, can, be allowed to* *Children have to go to school.* *I can stay at my sister's house.* *We're allowed to wear jeans.* p31 *should, must* *We should take traveller's cheques.* *You must write to us every week.* p33	Nationality words *Japan* *the Japanese* *Spain* *the Spanish* Countries and adjectives *Greece* *Greek* *Italy* *Italian* p36	Requests and offers *Could you … ?* *Would you … ?* *Can I … ?* *I'll …* *Shall I … ?* p37
5 On the move p38	**Future forms** *going to* and *will* *I'm going to buy some.* *I'll get a loaf.* p39 Present Continuous *We're playing tennis this afternoon.* p39	The weather *It's sunny.* *sunshine* *The sun's shining.* p44	Travelling around Using public transport Requests in a hotel p45
6 I just love it! p46	**Questions with *like*** *What's she like?* *What does she look like?* *What does she like doing?* p47 **Verb patterns** *I enjoyed meeting your friends.* *I just wanted to say thank you.* *You made me feel welcome.* p49	Describing food, cities, and people *fresh* *polluted* *sophisticated* p52 Collocations *fresh food* *historic cities* *elderly people* p52	Signs and sounds *Dry clean only* *Just looking, thanks.* p53
Stop and Check 2 Teacher's Book			

SKILLS DEVELOPMENT

READING	SPEAKING	LISTENING	WRITING p102
'Wonders of the modern world' – amazing technological and scientific achievements p10	Information gap – a UN Goodwill Ambassador p9 Discussion – what's the most important invention? p12	My wonders – three generations give their ideas about the wonders of the modern world p12	Correcting mistakes (1) – finding and correcting language mistakes in an informal letter p103
'The clown doctor' – a woman describes the job she loves p18	Discussion – what makes people happy? p14	Sports – three people talk about their free time activities p21	Letters and emails p104
'The painter and the writer' – the lives of Pablo Picasso and Ernest Hemingway (jigsaw) p26	Information gap – 'An amazing thing happened!' p25 Describing a book or a film you like p28	Books and films – people talk about their favourite books and films p28	A narrative (1) p106
'A world guide to good manners' – how to behave abroad p34	Talking about rules and regulations p32 Roleplay – starting a new job p33 Discussion – what advice would you give a foreign visitor? p34	Come round to my place! – entertaining friends in three different countries p36	For and against p108
'My kind of holiday' – a travel agent talks about his holidays p42	Arranging to meet p41 Discussion – your ideal holiday p42	A weather forecast p44	Making a reservation p109
'Global pizza' – the history of the world's favourite food p50	Talking about popular food and popular places to eat p50 Discussion – restaurants, cities and people you know p52	New York and London – An English couple talks about living in New York; an American gives his impressions of living in London (jigsaw) p52	A description (1) p110

Appendix 5.4. Learning outcomes of the Academic English Skills Course (level A)

A: Vocabulary
By the end of the course students will be able to do the following:

- Use a monolingual dictionary
- Use word cards to record vocabulary
- Construct antonym and synonym relationships
- Use diagrams to record vocabulary
- Use synonyms to avoid repetition in writing
- Identify parts of speech using affixes
- Recognise and use common abbreviations
- Use receptively and productively 1000 GSL, AWL and off-list words (see
- Headway Academic Skills 2 for specifics)

B: Reading

By the end of the course students should be able to do the following:

- Read an extensive text of around 1000 words broadly relevant to an area of
- study and respond to questions that require analytical skills, e.g. prediction,
- deduction, inference
- Name and describe the difference between the four principal reading strategies: skimming, scanning, extensive and intensive reading.
- Skim for gist/main points
- Scan for details
- Read closely for detailed understanding (intensive reading)
- Read longer texts e.g. graded readers (extensive reading)
- Read a range of text types: newspaper articles, general-audience technology magazine articles, programs, schedules, letters, forms with gist, main points and detailed comprehension
- Infer meanings of words from supportive contexts

- Make effective notes on a text such that the student can reconstruct the main points of the text.

    - Recall and define main concepts.
    - Use English rather than Arabic for notes in margins and glossing vocabulary.
    - Support key points with relevant additional details.
    - Organise information to enable quick reference at a later date.
    - Date notes.
    - Use notes to create a summary.
    - Sort out information and reject irrelevant pieces

- Recognise the hedging function of modals (e.g. may, could) and adverbs of possibility (e.g. possibly)
- Make inferences about information not stated in texts.
- Interpret graphic information
- Use a contents page and an index to locate information in a book.

C: Writing

By the end of the course students should be able to do the following:

- Produce a written report of a minimum of 500 words showing evidence
- of research, notetaking, review and revision of work, paraphrasing, summarising, use of quotations and use of references
- Proof-read effectively focusing on a range of surface features
- Complete applications forms
- Use mind-maps to brainstorm content for writing
- Use linking words to show logical organisation within and across sentences
- Reformulate phrases from a sentence
- Paraphrase sentences from a text
- Summarise paragraphs from a text
- Use pronouns to avoid repetition
- Use modal verbs (e.g. may, could) and adverbs of possibility (e.g. possibly)
- to hedge

- Transfer information from graph to text and text to graph.
- Cite sources according to the APA system
- Plan and execute a piece of writing by moving through a series of process stages

D: Listening

By the end of the course students should be able to do the following:

- Take notes on peer presentations, sufficient to enable the student to re-construct the main points of the presentation
- Take notes on longer talks/mini-lectures (10-15 minutes)
- Use prediction techniques to support understanding.
- Use discourse markers and lecturer signals to support understanding.
- Ask for clarification/repetition/rephrasing

E: Speaking

By the end of the course students should be able to do the following:

- Prepare and deliver a talk of at least 5 minutes. Use library resources in preparing the talk, speak clearly and confidently, make eye contact and use body language to support the delivery of ideas. Respond confidently to questions.

    - Outline and define main concepts.
    - Address questions from the audience.
    - Plan and conduct a presentation based on information from written material, interviews, surveys, etc.
    - Speak in a clearly audible and well paced voice.
    - Follow a presentation format.
    - Use presentation language (discourse markers etc.).
    - Achieve the key aim of informing the audience.
    - Make use of audio/visual aids when giving oral presentations.
    - Tailor content and language to the level of the audience.
    - Maintain some eye contact with audience.
    - Speak from notes in front of an audience using index cards.

- Observe time restrictions in presentations.
- Organise and present information in a logical order at a comprehensible speed.
- Invite constructive feedback and self-evaluate the presentation.

- Participate in a small group discussion using appropriate strategies to gain and concede the floor, make a point, interrupt, disagree

F: Other
By the end of the course students should be able to:
- Prepare an information search plan
- Use the LRC system for finding, borrowing and returning library material.
- Locate a book/journal in the library using the catalogue.
- Find specific information using internet search engines and electronic resources.
- Select or reject a source based on difficulty level, relevance and currency.
- Assess the reliability, objectivity and authenticity of a source.
- Create term planners and study schedules noting key dates/events.
- Organise a feasible study schedule that accommodates other responsibilities.
- Describe learning experiences, challenges, insights in a journal.
- Keep a portfolio of their work

LEARNING OUTCOMES

Learning outcomes are specified below by Level (C, B or A) and by course (General English Skills or Academic English Skills). Those outcomes in black are covered by the existing materials provided (*New Headway Plus* and *Academic Skills*). Those learning outcomes marked in blue are not covered by the existing materials and will need to be addressed independently by teachers. (It is anticipated that by the end of the 2010-11 academic year in-house material addressing these areas will have been created and collated.)

Appendix 5.5. Band descriptors used to evaluate the AES written project, retrieved from coordinators' materials website January 2011

Foundation Year: AES Written Project

Pass ***Pass*** ***Pass*** ***Fail*** ***Fail*** ***Fail***

	20 - 25	**16-20**	**10 - 15**	**6-9**	**1-5**	**0**
TASK ACHIEVEMENT	All outlines and drafts completed and submitted on time Student has actively tried to implement all changes suggested by teacher. Majority of the essay is in the students own words and credit is given when others work is used. Addresses chosen topic directly; coverage is fairly comprehensive; little irrelevance. Essay structure used includes introduction, conclusion, thesis statements and topic sentences. Meets minimum word limits	All outlines and drafts completed and submitted. Student has tried to implement most changes suggested by teacher. Majority of the essay is in the students own words and credit is given when others work is used. Addresses chosen topic but some points may not be covered or some irrelevance may appear Essay structure used includes introduction, conclusion and topic sentences. Meets minimum word limits	Most outlines and drafts were completed and submitted on time. Student has tried to implement some changes suggested by teacher. May contain a small amount of copied material. Some attempt to paraphrase. Addresses chosen topic but contains irrelevant points and some relevant points are not dealt with. Essay structure used includes recognisable introduction and conclusion. Meets minimum word limits	Some outlines and drafts were completed and submitted on time. Student has tried to implement changes suggested with limited success. May contain substantial amounts of copied material. Limited attempt to paraphrase. Limited relation to the chosen topic: shows some attempt to address the issue but contains little relevant material. No recognizable introduction and conclusion. May be short.	Most outlines and drafts were not submitted/not submitted on time. Student has made minimal/no attempt to implement changes and with little success. May contain mostly copied material. No formal essay structure used. Answer bears no or almost no relation to task.	Outlines and drafts not submitted/ not submitted on time. Student has made no attempt to implement changes. No assessable sample i.e. nothing legible/original in the essay.

Only give full marks i.e no half marks.

Appendix 5.6. Descriptors for assessing the student presentations in AES, retrieved from the coordinators' materials website in January 2011.

Speaking test assessment criteria: Foundation Levels A, B and C

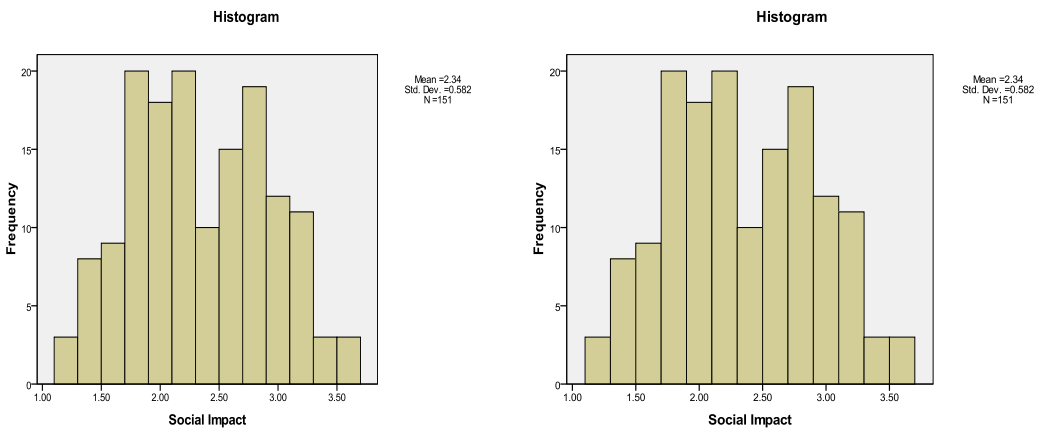
Score	Fluency	Grammar	Vocabulary	Phonology	Task achievement
5	Able to sustain flow of language necessary to accomplish the tasks with some pauses to search for words. No strain on listener.	Candidate has the range of grammar necessary to accomplish the tasks and is generally accurate.	Candidate has the range of vocabulary necessary to accomplish the tasks.	L1 features present but do not cause difficulty in understanding	Tasks accomplished fully and effectively
4	Able to sustain flow of language necessary to accomplish the tasks but with frequent pauses to search for words. Some strain on the listener.	Candidate either lacks the full range necessary to accomplish the tasks or has the range but lacks accuracy**.**	Candidate lacks the full range of vocabulary necessary but achieves communication through simple paraphrase	L1 features present and occasionally obstruct understanding	Tasks accomplished adequately
3	Pauses to search for words are so frequent that the flow of language necessary to accomplish the tasks is not sustained. Listener requires some patience.	Candidate lacks both the range and accuracy necessary to accomplish the tasks.	Candidate lacks the range of vocabulary necessary to accomplish the tasks and has limited ability to paraphrase.	L1 features strongly present and often obstruct understanding. Up to half of the speech is unintelligible.	Tasks accomplished to a limited degree
2	Speech frequently disconnected and very difficult to follow. Listener has to be pro-active to construct meaning.	Candidate has range and accuracy sufficient to attempt but not to accomplish the tasks.	Candidate has insufficient range to accomplish the tasks and cannot paraphrase.	L1 features strongly present and constantly obstruct understanding. Speech is largely unintelligible.	Tasks attempted but not accomplished
1	No connected speech	Range and accuracy inadequate for the test.	Range wholly inadequate for the test.	Speech unintelligible	Tasks not attempted
0	No adequate sample of language				

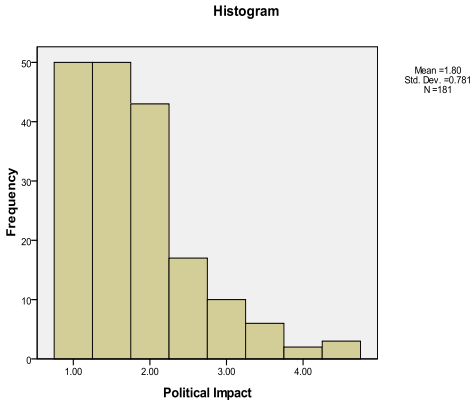
Appendix 6.1. Kolmogorov-Smirnov tests of normality for the student questionnaire in Phase 1[a]

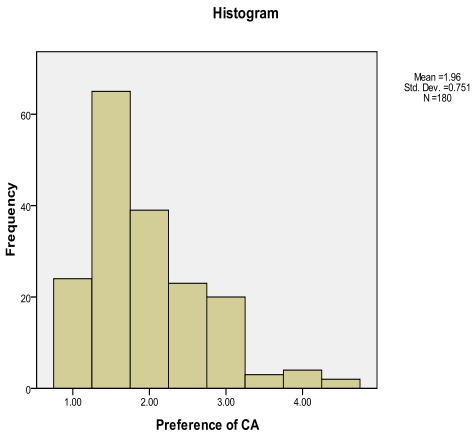
Topics	Kolmogorov-Smirnova			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Social Impact	.113	151	.000	.971	166	.001
Preference of CA	.224	180	.000	.916	178	.000
Preference of Test	.175	182	.000	.943	177	.000
Political Impact	.203	181	.000	.913	180	.000
Reliability	.222	180	.000	.984	146	.089
Construct Validity	.154	177	.000	.853	181	.000
Content Validity	.218	178	.000	Statistic	df	Sig.
Satisfaction	.157	166	.000	.971	166	.001

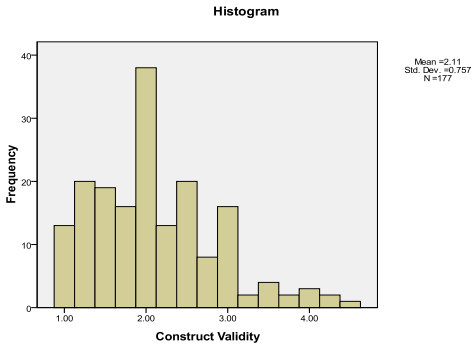
[a]For a normal distribution the **Sig.** values in this test should be > .05 (Pallant, 2007). Appendix 6.1 shows that the values of **Sig.** for the topics are all <.000.

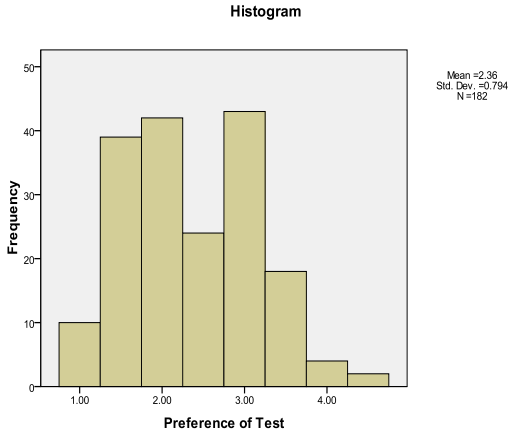
Appendix 6.2. Histograms of the students' responses to each topic in the Student Questionnaire in Phase 1

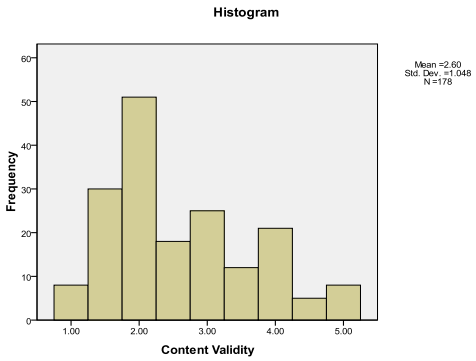


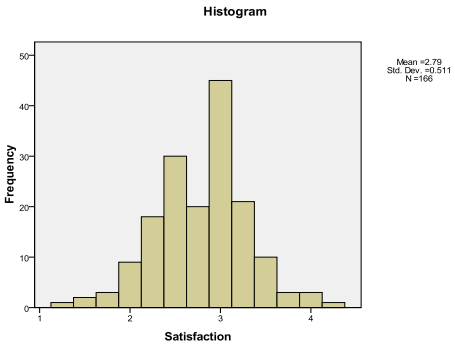

Histogram
Mean =1.80
Std. Dev. =0.781
N =181
Frequency
Political Impact

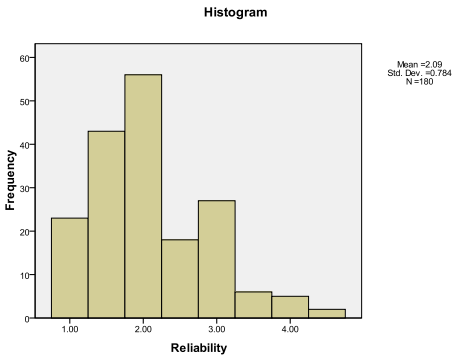

Histogram
Mean =1.96
Std. Dev. =0.751
N =180
Frequency
Preference of CA


Histogram
Mean =2.11
Std. Dev. =0.757
N =177
Frequency
Construct Validity


Histogram
Mean =2.36
Std. Dev. =0.794
N =182
Frequency
Preference of Test


Histogram
Mean =2.60
Std. Dev. =1.048
N =178
Frequency
Content Validity


Histogram
Mean =2.79
Std. Dev. =0.511
N =166
Frequency
Satisfaction

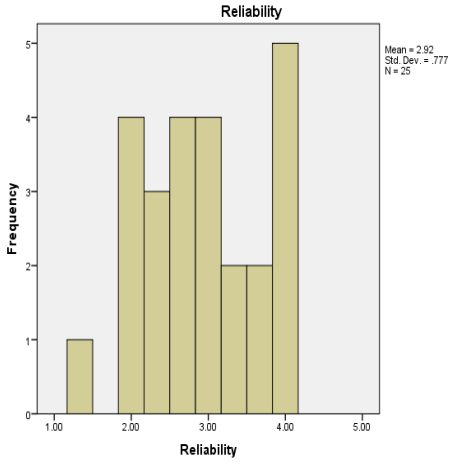

Histogram
Mean =2.09
Std. Dev. =0.784
N =180
Frequency
Reliability

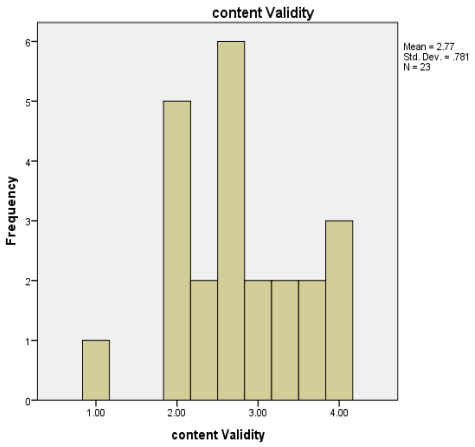
Appendix 6.3. Kolmogorov-Smirnov tests of normality for the teacher questionnaire in Phase 1[a]

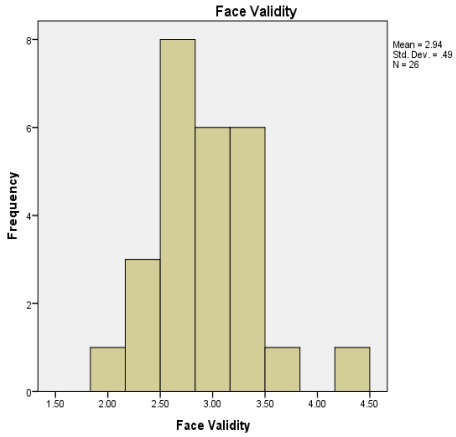
Topic	Kolmogorov-Smirnova			Shapiro-Wilk		
	Statistic	df	Sig.	.942	23	.199
Reliability	.118	25	.200*	.931	26	.083
Content Validity	.160	23	.129	.957	26	.341
Inappropriateness	.170	26	.051	.884	26	.007
Construct Validity	.119	26	.200*	.932	27	.079
Test/CA	.266	26	.000	.981	24	.906
Centrality of Assessment Writing	.147	27	.139	.916	27	.032
Marking Experience	.146	27	.147	.956	25	.340
Social Impact	.104	24	.200*	.954	27	.261
Political Impact	.170	27	.044	.942	23	.199
Predictive Validity	.173	25	.053	.931	26	.083

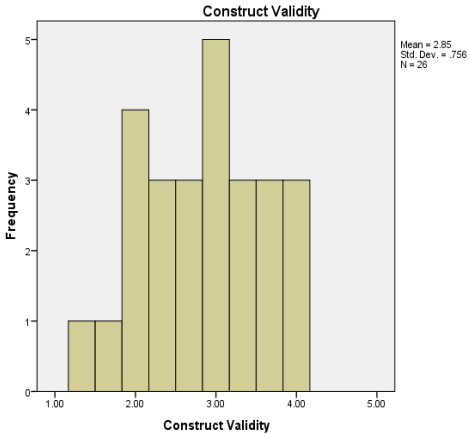
[a]According to Kolmogorov-Smirnov test of normality, all responses to the ten topics (see appendix 6.3 and 6.4) were fairly normally distributed but to four (i.e. *Face Validity, Test/CA, Political Impact* and *Predictive Validity*). A normal distribution should obtain a **Sig**. > 0.05 (Pallant, 2007).

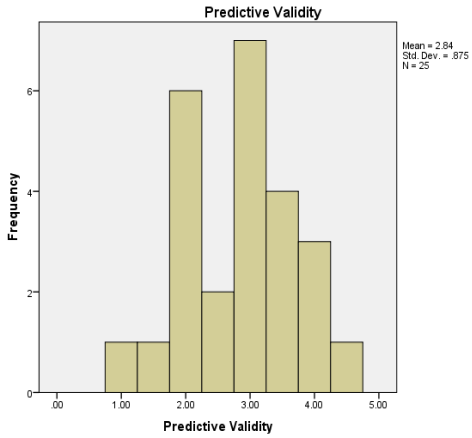
Appendix 6.4. Histograms of the teachers' responses to each topic in the teacher questionnaire in Phase 1

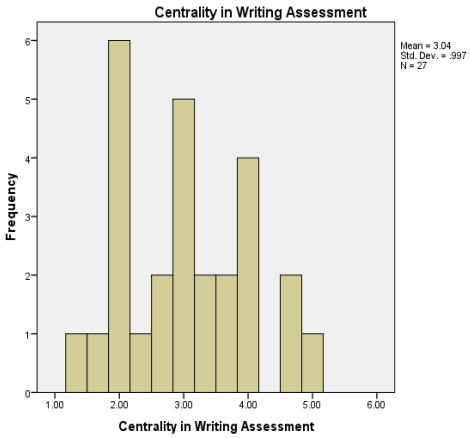


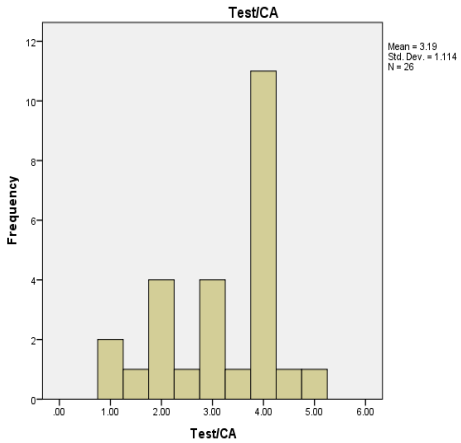


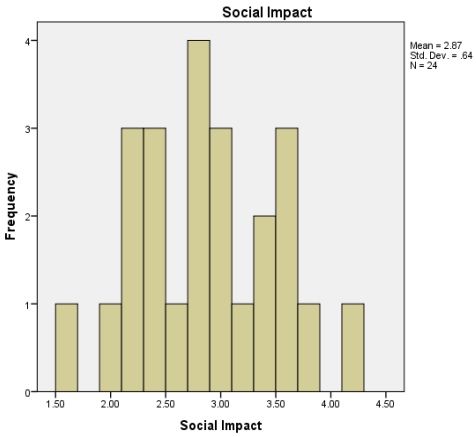

Face Validity
Mean = 2.94
Std. Dev. = .49
N = 26
Frequency
Face Validity

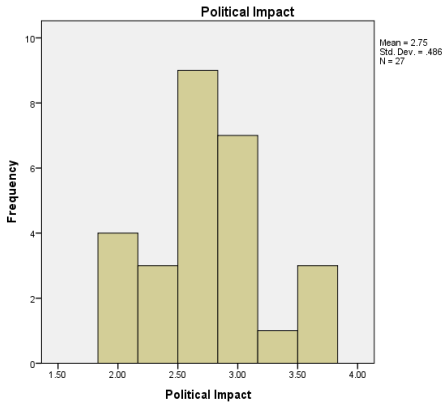

Construct Validity
Mean = 2.85
Std. Dev. = .756
N = 26
Frequency
Construct Validity


Predictive Validity
Mean = 2.84
Std. Dev. = .875
N = 25
Frequency
Predictive Validity


Centrality in Writing Assessment
Mean = 3.04
Std. Dev. = .997
N = 27
Frequency
Centrality in Writing Assessment


Test/CA
Mean = 3.19
Std. Dev. = 1.114
N = 26
Frequency
Test/CA


Social Impact
Mean = 2.87
Std. Dev. = .64
N = 24
Frequency
Social Impact


Political Impact
Mean = 2.75
Std. Dev. = .486
N = 27
Frequency
Political Impact


Confidence in Writing and Marking Tests
Mean = 2.48
Std. Dev. = .813
N = 27
Frequency
Confidence in Writing and Marking Tests

Appendix 7.1. Literal and edited translation of focus group 2 in phase1 (the first page of the transcribed focus group discussion only)

Literal Translation	Edited Translation
In this translation, the discussions, speaking turns and utterances closely resemble what happened in the focus group.	In this translation, the discussions were edited by omitting or rearranging interruptions, hesitations, or non-comprehensible input. This method of translation attempts to presents the data in a compressible and chronological manner without losing important features of the actual discussion.
Facilitator: How will you be assessed in the general and academic courses? Student6: in the general we are assessed by tests only and in the academic we are assessed by the essay and presentation. Student4: the first thing …there is…for the academic, there is 50% on the essay Student8: and presentation Student7: yeah Student4: and there is 50% on the presentation Student8: In the essay we use different resources…um..uhh.. from different places like internet, English book, Student5: yes and Arabic books. Student8: right and we do a presentation. Student4: from the report. Student6: no, no you cannot, listen..uhhm um.. Student8: the presentation is based on the report. Students7: but …um the teachers are not clear about the essays and what they want us to do…I mean ... we always change things Student4: yeah Student7: I don't feel that we gained anything. It is a waste of time … you know... to work on one thing and keep on changing it. I do not know. Student6: well…um ...I think that I think that allocating 50% in the academic for the essay is too much…you know … there should be more essays and the mark should be distributed on more activities and tasks. Student 5: yes, yes, I agree, the marks distribution should be changed, there is a textbook, I mean we study a textbook for no marks for doing it, for the activities; there are no tests or quizzes on it. Student2: I agree that…	**Facilitator: How will you be assessed in the general and academic courses?** Student6: In the academic English we are assessed by presentation and an essay while in the general we are assessed by tests only. Student4: 50% on the essay and 50% on the presentation in regard to the academic. Student 8: In the essay we use resources from different places like the internet and book written in English and the presentation is based on them. Student 7: The teachers are not clear about the essays and what they want us to do, we always change things in the essay. I do not feel that we gained anything. It is a waste of time to work on one thing and keep on changing it. Student6: I think that allocating 50% of the marks in the academic course to the essay is too much, there should be more essays and the mark should be distributed on more than one activity. Student 5: I agree that the marks distribution in the academic course should be altered, we have a textbook that we study in this course but there are no marks for studying it, there is no assessment on it. It is only the essay. So when the teacher explains what is in the book we do not listen, we only care about the essay. It is a waste of time. Student4: But there are some important things in the book like the vocabulary we get. Student 5: we study the textbooks in the general and we learn more from the general than the academic we learn reading writing grammar and vocabulary. But we do not learn anything from the academic; we just write the essay and do the presentation. Student7: I agree that we do not get anything

Student4: But… Students 5: yeah…I mean… It is only the essay. So when the teacher explains what is in the book we do not listen! Students: laughs Student5: we only care about the essay. It is a waste of time! Student4: yeah..but you know...But there are some important things in the book like the vocabulary we get. Student7: I agree … that we do not get anything from the book. in the academic course…. we only study the textbooks …in the general course and work on the essay on the academic course. Oh God…We have not even started working on the presentation yet! Can you imagine! Student8: ok…uhhh, you know.. we do not know how the essay will be marked ummm we just keep on changing and changing and changing in the essay and hope we get marks. Leave it to God, what could we do..huh Some of the groups have already submitted the essays but not our group. Student4: yeah.. Student 3: we do not know how we will be assessed in the presentation too. Student4: yes we do not know how we will be assessed ha(sigh). Student 8: hehehe (laugh)We have not started preparing for the presentation ..umm..when will it be guys? Student3: Next week Student 1: we prefer the tests more. Atudent8:oh yeah.. it will be next week and we do not know how we will be assessed…uumm and.. All of the past months we were studying how to write the essay only. Student7: Wait, ..umm…we, originally… prefer assessment last semester.. Student 5: no no Student7: listen…. Because had quizzes during the semester in the general course before doing the final. Student3: yeah, I agree…the focus should not be on the tests on the general course and essay on the academic, there should be a variety of assessment tools. Student 5: no …but .. you know..there is a variety in the general course the test tests on grammar, ummm reading um ..writing and listening unlike the academic course where	from the book we only study the textbooks on the general course and work on the essay on the academic course. We have not even started working on the presentation yet. Student8: we do not know how the essay will be marked, we just keep on changing the essay and hope we get marks. Some of the groups have already submitted the essays but not our group. Student 3: we do not know how we will be assessed in the presentation too. Student4: yes we do not know how we will be assessed. Student 8: We have not started preparing for the presentation which will be next week and we do not know how we will be assessed. All of the past months we were studying how to write the essay only. Student 1: we prefer the tests more. Student7: we prefer assessment last semester because we had quizzes during the semester in the general course before doing the final. Student3: The focus should not be on the tests on the general course and essay on the academic, there should be a variety of assessment tools. Student 5: there is a variety in the general course the test tests on grammar, reading writing and listening unlike the academic course where the focus is the essay. We need more activities than the essay and presentation. Student 8: The system of assessing all of the skills altogether is not suitable for us. We would like it to be focused on each skill. Grammar is tested separately from the reading and listening. This gives us more focus on studying each skill alone. Student7: Sometimes some teachers cannot teach grammar so when all of the language skills are combined into one course students suffer that they do not get the best teaching possible.

the focus is the essay uhhha We need more activities than the essay and presentation. Student 8: no …wait.. the system of assessing all of the skills altogether is not suitable for us.umm.. We would like it to be focused on each skill. Grammar is tested separately from the reading and listening…uhh.. Student7: But… Student 8: …, I think,, listen …I think that this gives us more focus on studying each skill alone. Student7: ok…but …you know that …sometimes some teachers cannot teach grammar ...uhh, so, uhhh when all of the language skills are combined into one course students suffer becuase they do not get the best teaching possible.	

Appendix 8.1. Kolmogorov-Smirnov and Shapiro-Wilk Test and Skewness Values for the responses to the student questionnaire in phase 2[a]

Topics	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Assessing English language and ideas in content and English courses	.097	176	.000	.982	176	.025
Construct validity	.117	176	.000	.964	176	.000
Adequacy of English language level for FY study	.102	175	.000	.978	175	.007
Predictive Validity	.195	176	.000	.868	176	.000
Dissatisfaction with FP assessment	.129	175	.000	.949	175	.000
Consequence and impact	.171	176	.000	.938	176	.000

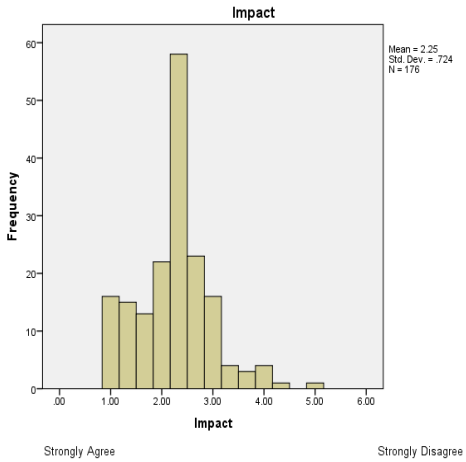
[a]A normal distribution is indicated by a non-significant result (Sig. value should be more than .05) in the Kolmogorov-Smirnov test. The values in this table reveal that the distribution of the responses to the *FY Construct Validity*, *FY Satisfaction* and *Social Impact* violated the assumption of normality.

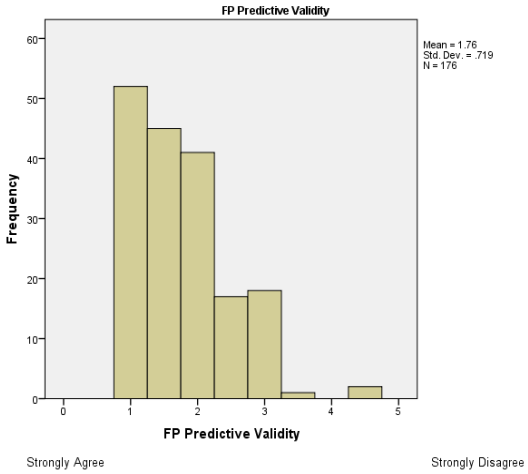
Skewness Values for the Student Questionnaire in Phase 2[a]

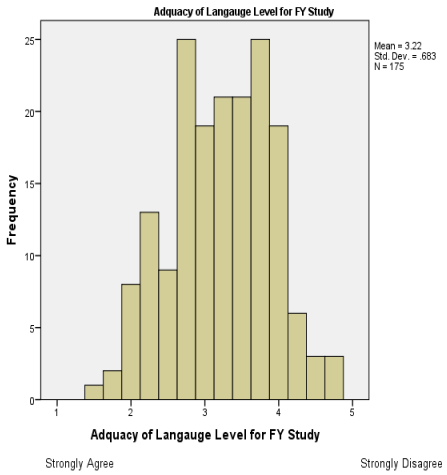
Topic	Skewness Values
Assessing English language and ideas in content and English courses	0.35
Construct validity	-.085
Adequacy of English language level for FY study	-0.12
Predictive validity	0.98
Dissatisfaction with FP assessment	0.18
Consequence and impact	0.42

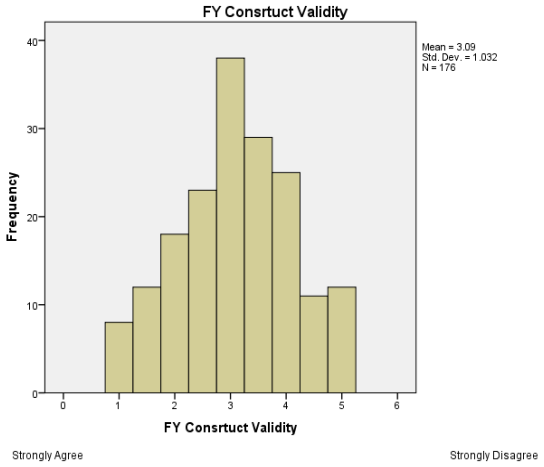
[a]The values of the skewness in appendix 8.3 show that the responses in all of the eight different topics were not normally distributed as a normal distribution will result in a skewness value of 0[a]

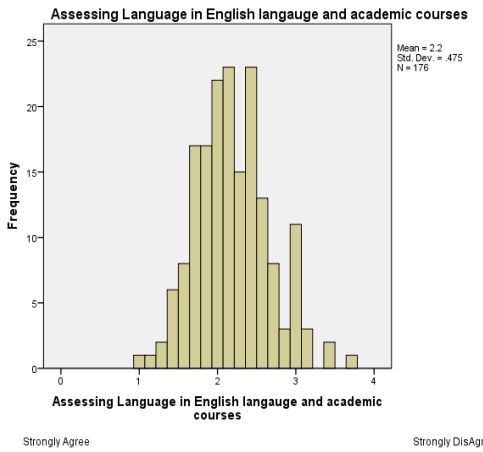
Appendix 8.2. Histograms of the responses to the student questionnaire in Phase 2

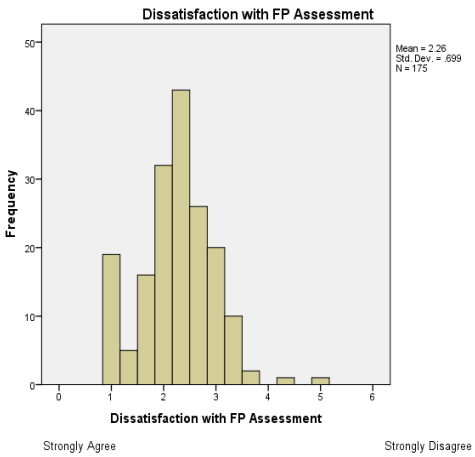










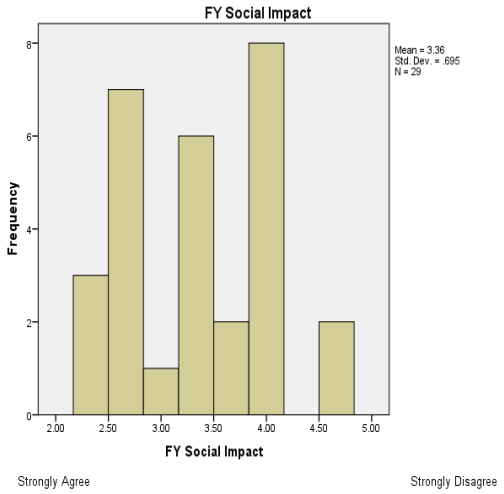


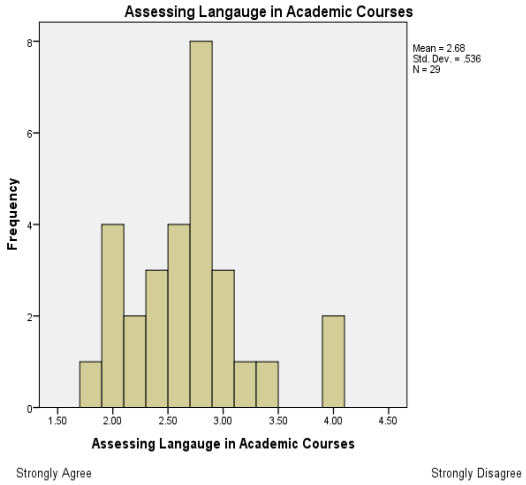
Appendix 8.3. Kolmogorov-Smirnov and Shapiro-Wilk Test and Skewness Values for the responses to the teacher questionnaire in phase 2

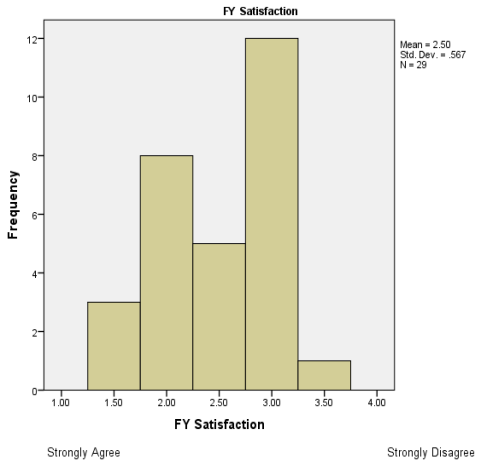
Topics	Kolmogorov-Smirnova			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Consistency in FY and FP assessment	.126	26	.200	.638	29	.000
Predictive Validity	.145	26	.168	.951	29	.190
Construct Validity	.231	26	.001	.952	29	.202
FP Satisfaction	.141	26	.199	.920	29	.030
FY Satisfaction	.262	26	.000	.963	29	.387
FY language criteria	.136	26	.200	.866	29	.002
Social Impact	.176	26	.037	.925	29	.042
Political Impact	.113	26	.200	.913	29	.021

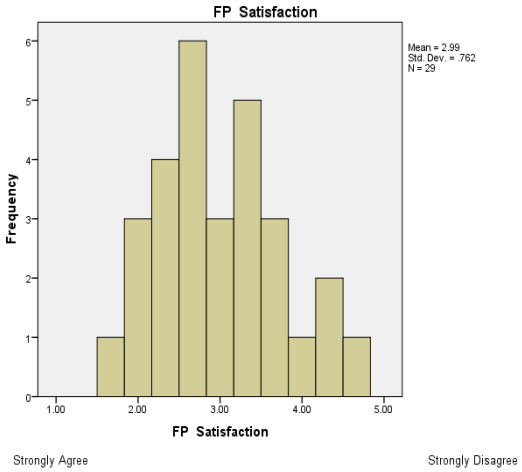
Topic	Skewness Values
Consistency in FY and FP	.513
Predictive Validity	-.385
Construct Validity	.132
Social Impact	.225
FY Satisfaction	-.385
Political Impact	.191
Assessing language in FY Academic Courses	.174
FP Satisfaction	.238

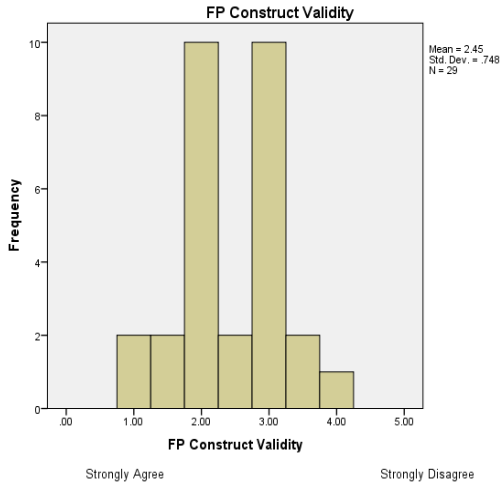
Appendix 8.4. Histograms of responses to the teacher questionnaire in phase 2

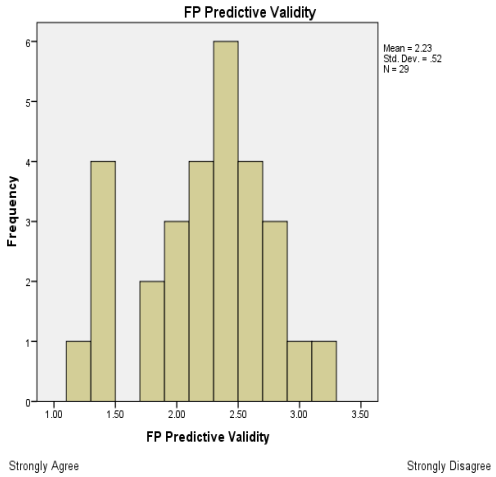


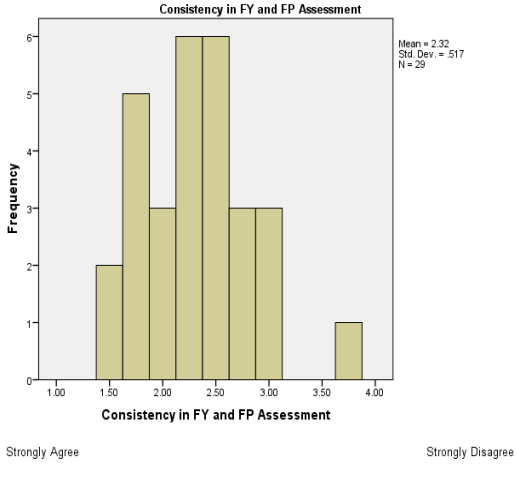


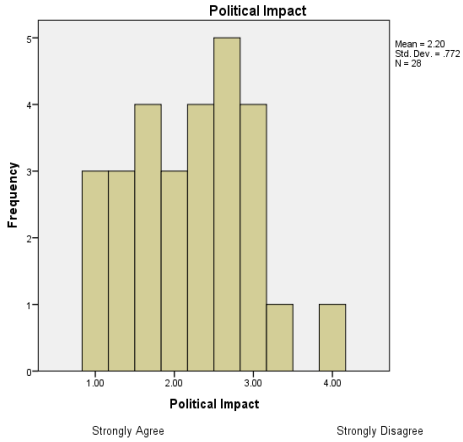





FP Construct Validity
Mean = 2.45
Std. Dev. = .748
N = 29
Frequency
FP Construct Validity
Strongly Agree
Strongly Disagree


FP Predictive Validity
Mean = 2.23
Std. Dev. = .52
N = 29
Frequency
FP Predictive Validity
Strongly Agree
Strongly Disagree

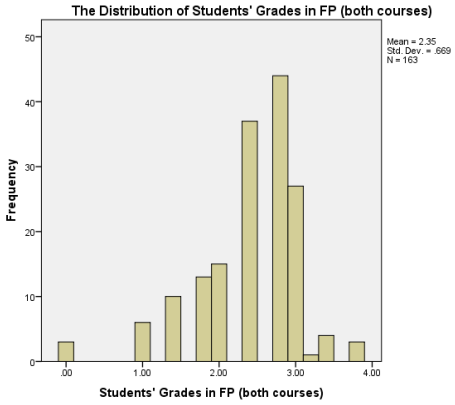

Consistency in FY and FP Assessment
Mean = 2.32
Std. Dev. = .517
N = 29
Frequency
Consistency in FY and FP Assessment
Strongly Agree
Strongly Disagree

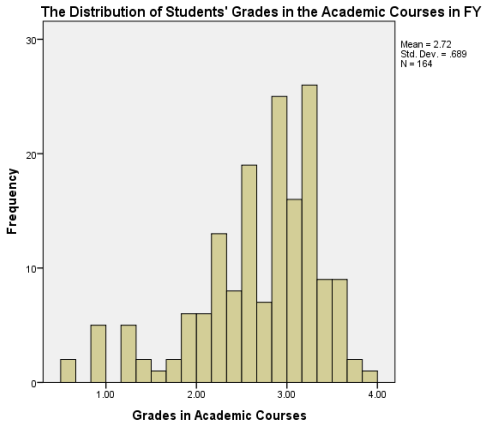

Political Impact
Mean = 2.20
Std. Dev. = .772
N = 28
Frequency
Political Impact
Strongly Agree
Strongly Disagree

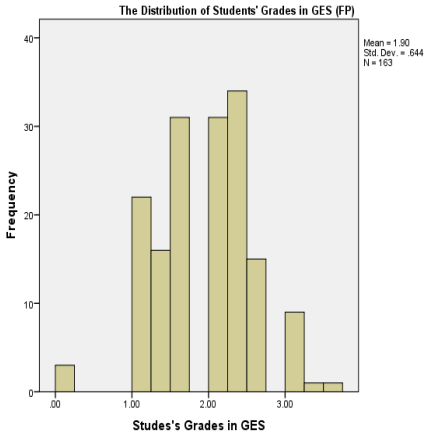
Appendix 10.1. Kolmogorov-Smirnov and Shapiro-Wilk Tests for the normality of distributions for the students' grades in academic courses, and FP courses (GES and AES).

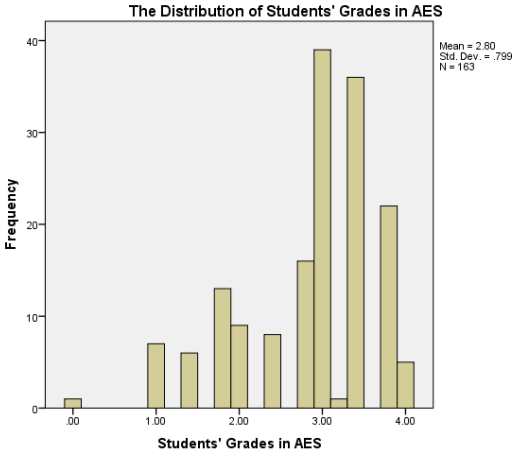
Courses	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Academic Courses	.129	164	.000	.919	164	.000
FP	.183	163	.000	.959	163	.000
GES	.125	163	.000	.899	163	.000
AES	.229	163	.000	.913	163	.000

Appendix 10.2. Histograms of students' grades in the academic courses, and FP courses (GES and AES).









- These are mid-term or final exam papers. Acronyms are as follows:
  SPR= Spring, AK= Answer Key, MT= Mid Term, SB= Student Book, V= version, D=draft, Aut=Autum, FN= Final, FA, Foundation level A,